Curs 2 - Fundamente de Big Data

Metode pentru predictie numerica - regresia

Gheorghe Cosmin Silaghi

Universitatea Babeș-Bolyai

March 14, 2023

Tipuri de *Analytics*

- *Analiza descriptiva*: furnizeaza o intelegere a ceea ce s-a intamplat
  - Profiling: caracterizeaza un comportament tipic al unui individ, grup sau populatie, prin generarea de statistici pe baza datelor
  - Sumarizare: Reduce cantitatea de informatie necesara pentru a directiona analiza pe aspectele critice
  - Clusterizare: impartirea indivizilor in grupuri intr-un mod folositor
- *Analiza Predictiva*: prezice (genereaza) informatie noua cu privire la instantele studiate
  - Clasificare: determina clasa unui instante
  - **Predictie numerica**: determina valoarea numerica exacta a unui atribut tinta al unei instante
- *Analiza prescriptiva*: cauta sa determine cea mai buna solutie dintre alternativele posibile la o problema

Tipuri de invatare

- **Supervizata**: sunt disponibile date cu rezultate cunoscute
- Nesupervizata: nu sunt disponibile cunostiinte anterioare cu privire la posibilele rezultate

Predictie

Predictie

Definitie

Fiind data o multime de *variabile independente*, sa se gaseasca un *model* care sa permita prezicerea unei *variabile dependente* (cu o acuratete cat mai mare posibila)

Predictie

Definitie

Fiind data o multime de *variabile independente*, sa se gaseasca un *model* care sa permita prezicerea unei *variabile dependente* (cu o acuratete cat mai mare posibila)

Tipuri de predictie

- *Clasificare*: variabila de iesire este nominala. Se prezice clasa unei instante
- *Predictie numerica*: se determina valoarea numerica exacta pentru variabila de iesire, asociata unei instante

Predictie

Definitie

Fiind data o multime de *variabile independente*, sa se gaseasca un *model* care sa permita prezicerea unei *variabile dependente* (cu o acuratete cat mai mare posibila)

Tipuri de predictie

- *Clasificare*: variabila de iesire este nominala. Se prezice clasa unei instante
- *Predictie numerica*: se determina valoarea numerica exacta pentru variabila de iesire, asociata unei instante

$$\text{Variabila dependenta} = f(\text{variabile independente})$$

Exemple de predictie numerica in business

- prezicerea ratei de parasire a firmei de catre angajati, in conditiile cunoasterii nivelelor salariale din domeniu, a starii globale a economiei, si a salariilor oferite de competitori
- la o banca, prezicerea pierderilor din credite acordate, in functie de maturitatea a creditelor, istoricul clientilor si gradul de colectare a ratelor
- estimarea ratei de refuz a aplicatiilor pentru credite pe baza ratei debit / venit, vechimea in munca a aplicantului, scorul de risc pentru creditare etc.

Predictie numerica - formalizare matematica

Y: variabila de iesire (dependenta) care trebuie prezisa
$X_1, X_2, \ldots, X_p$: p variabilele independente (predictori)

Problema: trebuie sa descoperim o relatie f intre Y si $X = (X_1, X_2, \ldots, X_p)$:

$$Y = f(X) + \epsilon \tag{1}$$

f: o functie necunoscuta, dar fixata, reprezinta *informatia sistematica* pe care X o furnizeaza pentru a prezice pe Y
ϵ: este *termenul de eroare*, aleator, de medie zero

De ce trebuie sa estimam f?

De ce trebuie sa estimam f?

Pentru predictie

- considerand ca $\hat{f}$ este o estimare pentru f, pentru un esantion X, putem prezice valorile $\hat{Y} = \hat{f}(X)$?
- acuratetea $\hat{Y}$ depinde de cat de bine l-am estimat (invatat) pe $\hat{f}$.
- *Eroarea reductibila*: cea care provine din eroarea de estimare a lui f prin $\hat{f}$
- *Eroare ireductibila*: cea datorata lui ϵ
- Tinta de invatare este sa il estimam pe $\hat{f}$ cat mai bine posibil, adica reducand eroarea de invatare $E(Y - \hat{Y})^2$ la varianta erorii ireductibile $Var(\epsilon)$

De ce trebuie sa estimam f?

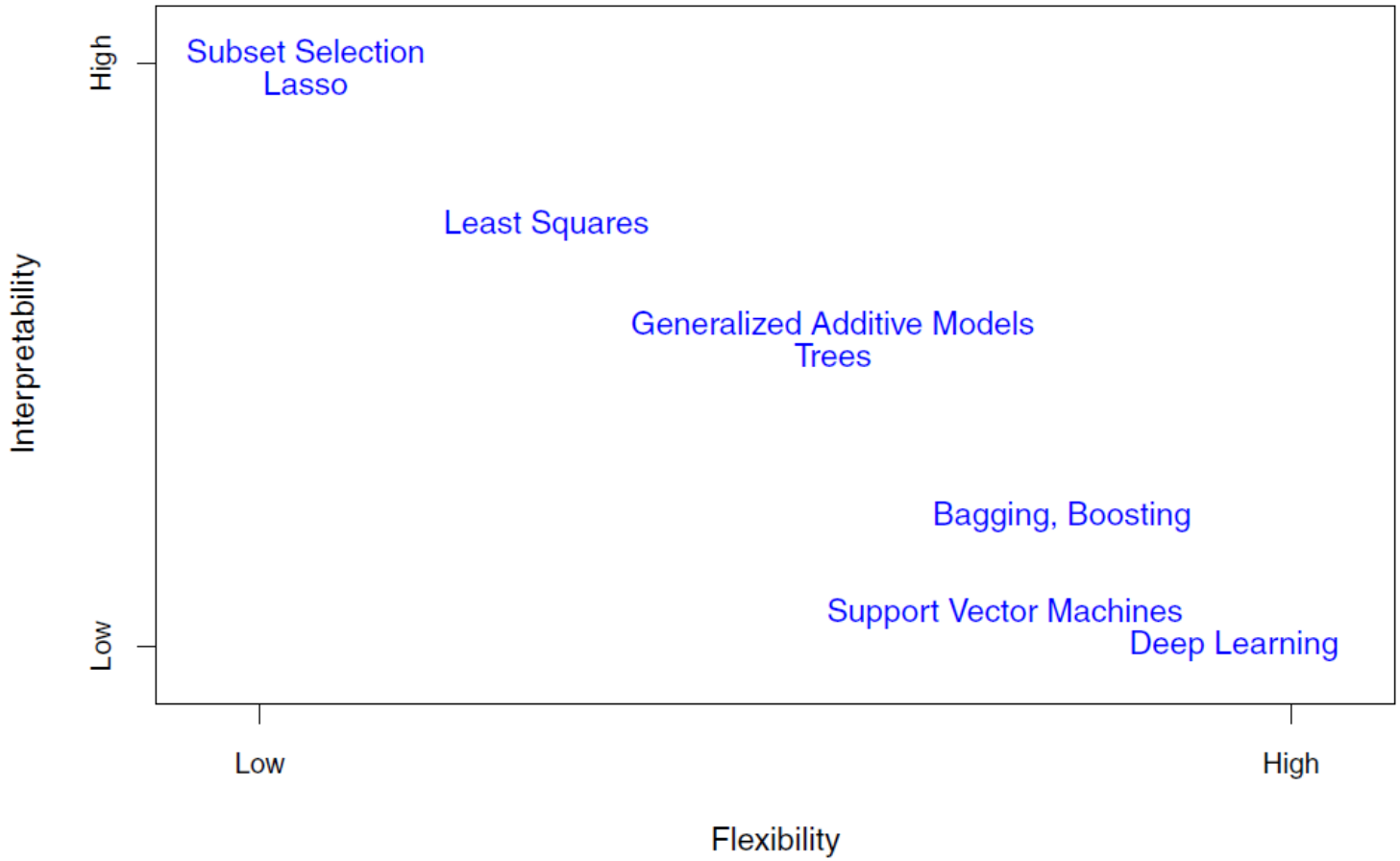
Pentru predictie

- considerand ca $\hat{f}$ este o estimare pentru f, pentru un esantion X, putem prezice valorile $\hat{Y} = \hat{f}(X)$?
- acuratetea $\hat{Y}$ depinde de cat de bine l-am estimat (invatat) pe $\hat{f}$.
- *Eroarea reductibila*: cea care provine din eroarea de estimare a lui f prin $\hat{f}$
- *Eroare ireductibila*: cea datorata lui ϵ
- Tinta de invatare este sa il estimam pe $\hat{f}$ cat mai bine posibil, adica reducand eroarea de invatare $E(Y - \hat{Y})^2$ la varianta erorii ireductibile $Var(\epsilon)$

Pentru inferenta

- dorim sa intelegem modul in care se schimba Y atunci cand unul sau mai multi predictori X_i se schimba
- care dintre predictori sunt (in mod relevant) asociati cu Y?
- care este relatia dintre variabila dependenta si fiecare dintre predictori?
- poate fi descrisa relatia dintre Y si predictori folosind o ecuatie liniara, sau este o relatie mai complicata?

Compromisul (trade-off) dintre acuratetea de predictie (flexibilitate) si interpretabilitatea (transparenta) modelului



G. James et al. An Introduction to Statistical Learning, Springer, 2021

Evaluarea acuratetii modelelor

De ce este necesar sa avem un numar asa mare de metode de invatare, si nu avem o singura metoda (cea mai buna)?

- Pentru ca niciuna dintre metode nu este mai puternica decat celelalte pe toate seturile de date !!!!!
- $\Rightarrow$ Pentru fiecare set de date (problema concreta), trebuie sa decidem care metoda de invatare este mai potrivita (in functie de intrebarea de cercetare specifica problemei)
- Pentru ca nu intotdeauna acuratetea de predictie este raspunsul dorit la o problema

Evaluarea acuratetii modelelor

De ce este necesar sa avem un numar asa mare de metode de invatare, si nu avem o singura metoda (cea mai buna)?

- Pentru ca niciuna dintre metode nu este mai puternica decat celelalte pe toate seturile de date !!!!!
- $\Rightarrow$ Pentru fiecare set de date (problema concreta), trebuie sa decidem care metoda de invatare este mai potrivita (in functie de intrebarea de cercetare specifica problemei)
- Pentru ca nu intotdeauna acuratetea de predictie este raspunsul dorit la o problema

Despre ChatGPT si esenta inteligentei artificiale

Noam Chomsky, The False Promise of ChatGPT, New York Times, 8 March 2023
Their deepest flaw is the absence of the most critical capacity of any intelligence: to say not only what is the case, what was the case and what will be the case — that's description and prediction — but also what is not the case and what could and could not be the case. Those are the ingredients of explanation, the mark of true intelligence.

Cum putem sa il estimam pe f?

Cum putem sa il estimam pe f?

Metode parametrice

1. facem o presupunere asupra formei lui f, de exemplu f este liniara (regresie):

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \tag{2}$$

2. invatam (sau potrivim) modelul selectat pe datele de antrenament.
3. ca si rezultat vom obtine estimari pentru parametrii $\beta_0, \beta_1, \ldots, \beta_p$

Cum putem sa il estimam pe f?

Metode parametrice

1. facem o presupunere asupra formei lui f, de exemplu f este liniara (regresie):

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \tag{2}$$

2. invatam (sau potrivim) modelul selectat pe datele de antrenament.
3. ca si rezultat vom obtine estimari pentru parametrii $\beta_0, \beta_1, \ldots, \beta_p$

Metode non-parametrice

- evitam sa facem o presupunere explicita asupra formei functionale a lui f
- incercam sa estimam pe f astfel incat sa ne apropiem pe cat posibil de datele de intrare,dar sa pastram in acelasi timp capabilitatea de generalizare
- de exemplu metode bazate pe retele neuronale (deep learning)

Masurarea calitatii procesului de invatare a lui f

Eroarea medie patratica - MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{3}$$

- MSE va fi mica daca raspunsurile prezise for fi apropiate de raspunsurile reale
- *training MSE* (MSE de invatare): este calculata pe datele folosite pentru invatarea parametrilor modelului (training data)
- *test MSE*: este calculata pe date noi, care nu au fost folosite in timpul procesului de invatare
- selectam modelul care ne da cel mai mic *test MSE*
- un model se spune ca face *overfitting*: daca training MSE este mic, dar test MSE este mare (modelul nu are capacitate de generalizare)

  $\rightarrow$ trebuie sa recunoastem si evitam aceste modele

Determinarea bugetelor de publicitate

- Setul de date prezinta vanzarile unui produs anume in functie de bugetele de publicitate cheltuite: pentru TV, radio si presa scrisa
- Fiecare instanta reprezinta un alt oras unde se vinde produsul

Intrebari de cercetare adresate:

1. Exista vreo relatie dintre bugetele de publicitate si vanzari?
2. Daca relatia exista, cat de puternica este?
   - → Fiind dat un buget de publicitate, putem prezice nivelul vanzarilor, si cu ce acuratete?
3. Care dintre canalele media contribuie la cresterea vanzarilor?
4. Cu ce acuratete putem estima efectul fiecarui mediu de publicitate asupra vanzarilor?
5. Cu ce acuratete putem prezice vanzarile viitoare?
6. Este relatia liniara?
7. Exista o sinergie intre mediile de publicitate? (efecte de interactiune)
   - → Fiind dat un buget de publicitate, putem imparti bugetul intre mediile de publicitate, sau este recomandat sa il cheltuim pe un singur mediu?

Setul de date si codul sursa: vezi moodle

Primii pasi

Primii pasi

1. Incarcarea datelor din csv in mediul de lucru (R)

Primii pasi

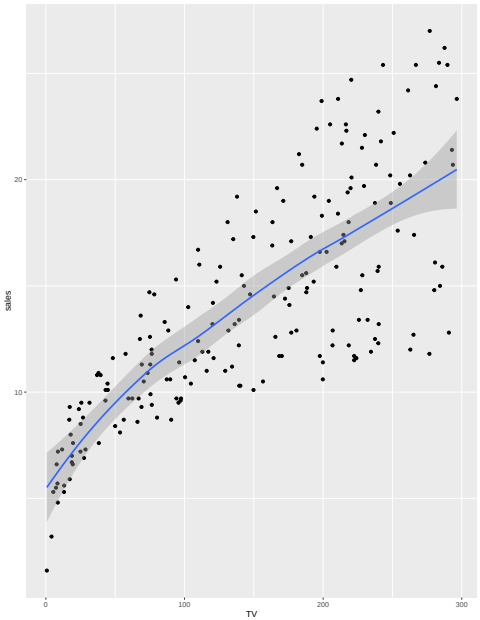
1. Incarcarea datelor din csv in mediul de lucru (R)
2. Explorarea tidy data si observarea atributelor nefolositoare

Primii pasi

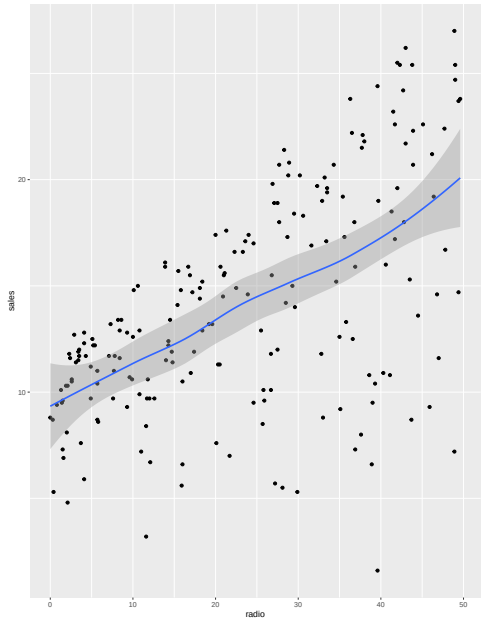
1. Incarcarea datelor din csv in mediul de lucru (R)
2. Explorarea tidy data si observarea atributelor nefolositoare
3. Realizarea de vizualizari a datelor pentru a intui tipul de relatie care ar putea exista intre predictori si variabila dependenta

Primii pasi

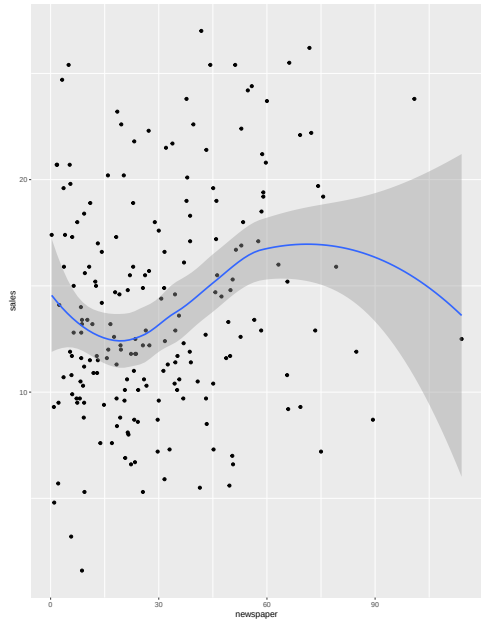
1. Incarcarea datelor din csv in mediul de lucru (R)
2. Explorarea tidy data si observarea atributelor nefolositoare
3. Realizarea de vizualizari a datelor pentru a intui tipul de relatie care ar putea exista intre predictori si variabila dependenta



(a) Vanzari vs. TV



(b) Vanzari vs. radio



(c) Vanzari vs. presa scrisa

Regresie liniara simpla

Regresie liniara simpla

Vanzarile in functie de bugetul pentru publicitate la TV

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV} \tag{4}$$

- β_0: intercept
- β_1: panta (slope)

Regresie liniara simpla

Vanzarile in functie de bugetul pentru publicitate la TV

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV} \tag{4}$$

- β_0: intercept
- β_1: panta (slope)

Cod R

```
mod_sales_TV <- lm(data = Advertising, sales ~ TV)
summary(mod_sales_TV)
```

Regresie liniara simpla

Vanzarile in functie de bugetul pentru publicitate la TV

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV} \tag{4}$$

- β_0: intercept
- β_1: panta (slope)

Cod R

```
mod_sales_TV <- lm(data = Advertising, sales ~ TV)
summary(mod_sales_TV)
```

Rezultate

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Interpretarea rezultatelor

- Std error (SE): ne spune (in medie) cu cat estimarea parametrului respectiv difera de valoarea reala a acestui parametru. Se folosesc pentru determinarea intervalelor de incredere pentru parametru
- t-statistic: $\frac{\hat{\beta}_i - 0}{SE(\hat{\beta}_i)}$ ne spune numarul de deviatii standard cu care estimarea parametrului β_i se departeaza de zero.
- p-value: indica probabilitatea ca sa observam o asociere intre predictor si variabila dependenta datorita sansei, in absenta unei asocieri reale intre predictor si variabila dependenta. O valoare mica a lui p-value ne permite sa tragem concluzia ca exista o asociere intre predictor si variabila dependenta.

Intervale de incredere pentru parametri

Erorile standard sunt folosite pentru a calcula intervalele de incredere pentru parametri β_i, cu o incredere de 95%

$$\beta_i \in [\hat{\beta}_i - 2SE(\hat{\beta}_i), \hat{\beta}_i + 2SE(\hat{\beta}_i)] \tag{5}$$

Intervale de incredere pentru parametri

Erorile standard sunt folosite pentru a calcula intervalele de incredere pentru parametri β_i, cu o incredere de 95%

$$\beta_i \in [\hat{\beta}_i - 2SE(\hat{\beta}_i), \hat{\beta}_i + 2SE(\hat{\beta}_i)] \tag{5}$$

- CI pentru β_0: $[6.13, 7.935]$: in absenta publicitatii la TV, vanzarile se vor situa in intervalul calculat
- CI pentru β_1: $[0.042, 0.053]$: o crestere de \$1000 a publicitatii TV va determina o crestere a vanzarilor intre \$42 si \$53

Intervale de incredere pentru parametri

Erorile standard sunt folosite pentru a calcula intervalele de incredere pentru parametri β_i, cu o incredere de 95%

$$\beta_i \in [\hat{\beta}_i - 2SE(\hat{\beta}_i), \hat{\beta}_i + 2SE(\hat{\beta}_i)] \tag{5}$$

- CI pentru β_0: $[6.13, 7.935]$: in absenta publicitatii la TV, vanzarile se vor situa in intervalul calculat
- CI pentru β_1: $[0.042, 0.053]$: o crestere de \$1000 a publicitatii TV va determina o crestere a vanzarilor intre \$42 si \$53

Testarea ipotezelor

H_0 Ipoteza nula: Nu se identifica nicio relatie intre X si Y: testam $H_0 : \beta_1 = 0$ versus

H_a Exista o relatie intre X si Y: si testam $H_a : \beta_1 \neq 0$

valori mici ale lui p-values indica faptul ca rejectam ipoteza nula respectiva

Evaluarea acuratetii modelului

- Eroarea standard reziduala (RSE): marimea medie cu care variabila dependenta Y va devia de la linia de regresie - masoara lipsa de potrivire (*lack of fit*) a modelului
- R^2: proportia din variabilitatea lui Y care poate fi explicata pe baza lui X. $R^2 \in [0, 1]$

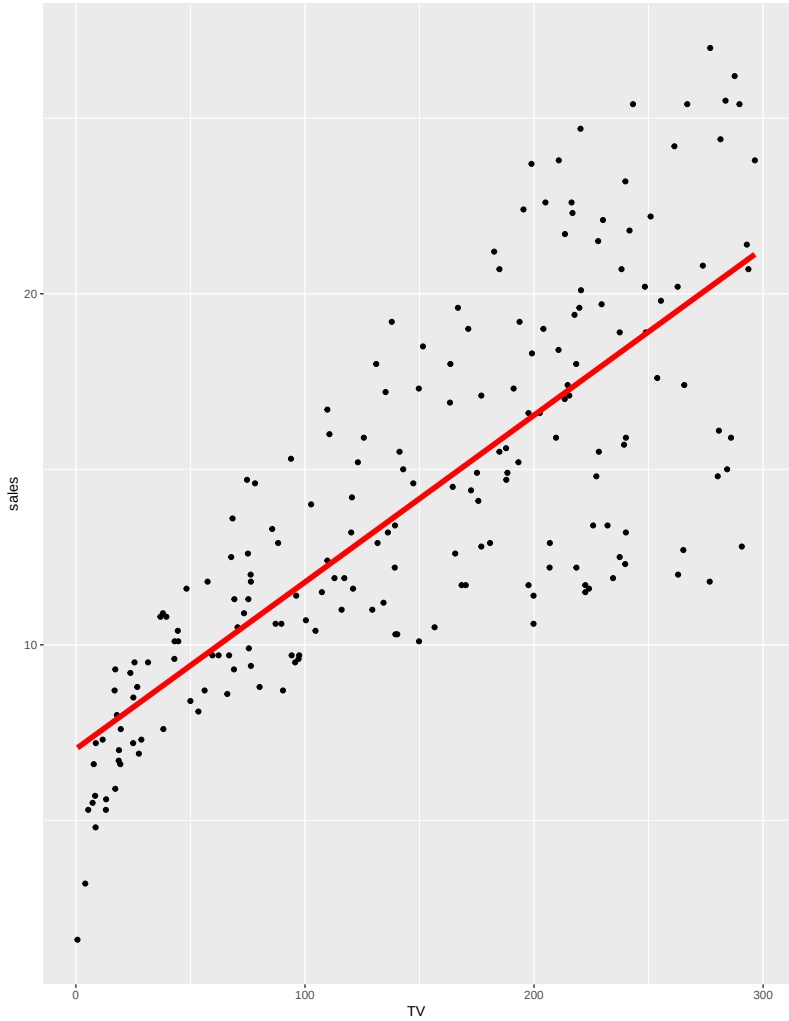
Masura	Valoare
RSE	3.26
R^2	0.612

Desenarea datelor si a liniei de regresie

Cod R

```
grid <- Advertising %>%
data_grid(TV = seq_range(TV, 100)) %>%
add_predictions(mod_sales_TV, "sales")

ggplot(Advertising, aes(TV, sales)) +
geom_point() +
geom_line(data=grid, color="red", size=2)
```



Regresie liniara cu mai multi predictori

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} \tag{6}$$

Regresie liniara cu mai multi predictori

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} \tag{6}$$

Cod R

```
mod_sales_all <- lm(data = Advertising, sales ~ TV + radio + newspaper)
summary(mod_sales_all)
```

Regresie liniara cu mai multi predictori

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} \quad (6)$$

Cod R

```
mod_sales_all <- lm(data = Advertising, sales ~ TV + radio + newspaper)
summary(mod_sales_all)
```

Rezultate

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Regresie liniara cu mai multi predictori

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} \quad (6)$$

Cod R

```
mod_sales_all <- lm(data = Advertising, sales ~ TV + radio + newspaper)
summary(mod_sales_all)
```

Rezultate

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Interpretare

Pentru un **nivel fixat** al publicitatii la TV si media scrisa, cresterea cu \$1000 a bugetului de publicitate radio determina o crestere a vanzarilor cu \$189.

Consideratii legate de interpretarea rezultatelor

- regresia simpla ne da o concluzie asupra relatiei dintre un predictor si variabila dependenta, in conditiile *ignorarii* celorlalti factori
- regresia multipla ne da o concluzie asupra relatiei dintre un predictor si variabila dependenta, in conditiile *fixarii nivelului* celorlalti factori
- factorul newspaper este slab relevant in regresia simpla
- factorul newspaper devine irelevant in regresia multipla

Intrebari adresate de regresia liniara multipla

1. Este cel putin unul din predictorii X_i folositori pentru a prezice variabila dependenta?
2. Toti predictorii ajuta la explicarea lui Y, sau doar un subset al acestora este folositor?
3. Cat de bine potriveste modelul pe datele disponibile?
4. Daca se da o valoare pentru fiecare predictor, putem prezice valoarea variabilei dependente? Care este acuratetea predictiei?

Relatia dintre predictori si variabila dependenta

Ipoteze

H_0 Ipoteza nula: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_a Ipoteza alternativa: H_a: at least one of β_j is not zero

- H_0 este acceptata daca *F statistics* se apropie de 1
- H_a este acceptata *F statistics* este (mult) mai mare decat 1

Masura	Value	p-value
RSE	1.686	
R^2	0.8972	
F statistics	570.3	< 0.0001

Relatia dintre predictori si variabila dependenta

Ipoteze

H_0 Ipoteza nula: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_a Ipoteza alternativa: H_a: at least one of β_j is not zero

- H_0 este acceptata daca *F statistics* se apropie de 1
- H_a este acceptata *F statistics* este (mult) mai mare decat 1

Masura	Value	p-value
RSE	1.686	
R^2	0.8972	
F statistics	570.3	< 0.0001

Mesaj important

Trebuie sa analizam atat valorile individuale ale lui t-statistics pentru fiecare predictor cat si F-statistics pentru intregul model (cu valorile p-values asociate) astfel incat sa rejectam in mod corect ipoteza nula H_0

Cum decidem care sunt variabilele importante?

- doar un subset de predictori sunt cu adevarat importanti - in special cand avem un numar mare de predictori p

$\Rightarrow$ putem realiza selectia atributelor (feature selection)

Cum decidem care sunt variabilele importante?

- doar un subset de predictori sunt cu adevarat importanti - in special cand avem un numar mare de predictori p

$\Rightarrow$ putem realiza selectia atributelor (feature selection)

Selectie inainte (forward)

1. calculam p regresii liniare simple si o consideram pe cea cu cel mai mic RSE
2. pornind de la modelul calculat la pct. 1, in mod succesiv calculam p-1 regresii prin adaugarea unei noi variabile, si o consideram pe cea cu cel mai mic RSE
3. continuam generarea regresiilor pana cand se indeplineste o conditie de oprire

Cum decidem care sunt variabilele importante?

- doar un subset de predictori sunt cu adevarat importanti - in special cand avem un numar mare de predictori p

$\Rightarrow$ putem realiza selectia atributelor (feature selection)

Selectie inainte (forward)

1. calculam p regresii liniare simple si o consideram pe cea cu cel mai mic RSE
2. pornind de la modelul calculat la pct. 1, in mod succesiv calculam p-1 regresii prin adaugarea unei noi variabile, si o consideram pe cea cu cel mai mic RSE
3. continuam generarea regresiilor pana cand se indeplineste o conditie de oprire

Selectie inapoi (backward)

1. incepem cu regresia multipla care contine toti cei p predictori
2. stergem pe rand predictorul cu cea mai mare valoare a lui p-value si calculam un nou model de regresie
3. continuam generarea regresiilor pana cand se indeplineste o conditie de oprire

Cum decidem care sunt variabilele importante?

- doar un subset de predictori sunt cu adevarat importanti - in special cand avem un numar mare de predictori p

$\Rightarrow$ putem realiza selectia atributelor (feature selection)

Selectie inainte (forward)

1. calculam p regresii liniare simple si o consideram pe cea cu cel mai mic RSE
2. pornind de la modelul calculat la pct. 1, in mod succesiv calculam p-1 regresii prin adaugarea unei noi variabile, si o consideram pe cea cu cel mai mic RSE
3. continuam generarea regresiilor pana cand se indeplineste o conditie de oprire

Selectie inapoi (backward)

1. incepem cu regresia multipla care contine toti cei p predictori
2. stergem pe rand predictorul cu cea mai mare valoare a lui p-value si calculam un nou model de regresie
3. continuam generarea regresiilor pana cand se indeplineste o conditie de oprire

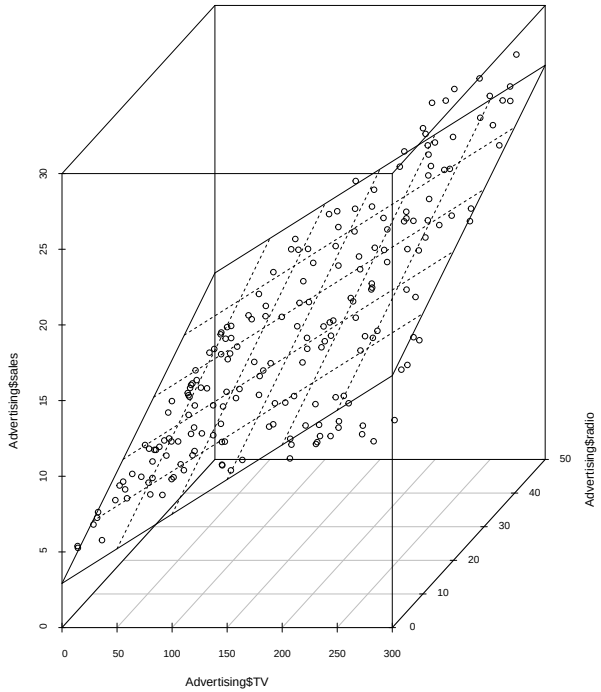
Selectie mixta (mixed)

realizam o combinatie a selectiei inainte cu selectia inapoi

Acuratetea potrivirii modelelor

Model	R^2	**RSE**
Vanzari vs. TV, radio, presa scrisa	0.8972	1.686
Vanzari vs. TV si radio	0.8972	1.681
Vanzari vs. TV	0.6119	3.259

- pe baza R^2 si RSE, decidem ca modelul cu TV si radio este cel cu acuratete mai buna
- putem imbunatati acest model adaugand un termen de interactiune intre cheltuielile pentru publicitatea la TV si radio



Vanzari vs.TV si radio - rezultate

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.921	0.2945	9.919	< 0.0001
TV	0.046	0.0014	32.91	< 0.0001
radio	0.188	0.0080	23.38	< 0.0001

Realizarea predictiilor (I)

Realizarea predictiilor (I)

Cod R

```
newspendings <- tibble( TV = 100,
radio = 20 )
predict(mod_sales_TV_radio, newdata = newspendings, interval = "confidence")
predict(mod_sales_TV_radio, newdata = newspendings, interval = "prediction")
```

Realizarea predictiilor (I)

Cod R

```
newspendings <- tibble( TV = 100,
radio = 20 )
predict(mod_sales_TV_radio, newdata = newspendings, interval = "confidence")
predict(mod_sales_TV_radio, newdata = newspendings, interval = "prediction")
```

Rezultate

	Potrivire	**Limita inferioara**	**Limita superioara**
Incredere	11.2564	10.9852	11.5277
Predictie	11.2564	7.9296	14.5833

Realizarea predictiilor (I)

Cod R

```
newspendings <- tibble( TV = 100,
radio = 20 )
predict(mod_sales_TV_radio, newdata = newspendings, interval = "confidence")
predict(mod_sales_TV_radio, newdata = newspendings, interval = "prediction")
```

Rezultate

	Potrivire	**Limita inferioara**	**Limita superioara**
Incredere	11.2564	10.9852	11.5277
Predictie	11.2564	7.9296	14.5833

- $\hat{\beta}_i$ sunt estimari pentru β_i reali, cu deviere datorata erorilor reductibile.

$\rightarrow$ calculam *intervalele de incredere* pentru a determina cat de apropiata este predictia $\hat{Y}$ de $f(X)$: se determina incertitudinea vanzarilor medii

- chiar daca cunoastem $f(X)$ real (sau mai precis toate valorile β_i), nu putem prezice precis, deoarece avem erorile ireductibile (datorate lui ϵ).

$\rightarrow$ calculam *intervalul de predictie* pentru a determina cat de aproape este $\hat{Y}$ de valoarea reala Y: se determina incertitudinea din jurul unei anume instante de test

Realizarea predictiilor (II)

Realizarea predictiilor (II)

- intervalul de incredere masoara incertitudinea existenta asupra valorii medii a variabilei dependente

$\Rightarrow$ daca *in fiecare* oras se cheltuie \$100000 pe publicitate TV si \$20000 pe publicitate radio, cu o incredere de 95%, valoarea (medie) a vanzarilor va fi situata intre \$10985.2 si \$11527.7

Realizarea predictiilor (II)

- intervalul de incredere masoara incertitudinea existenta asupra valorii medii a variabilei dependente

$\Rightarrow$ daca *in fiecare* oras se cheltuie \$100000 pe publicitate TV si \$20000 pe publicitate radio, cu o incredere de 95%, valoarea (medie) a vanzarilor va fi situata intre \$10985.2 si \$11527.7

- intervalul de predictie masoara incertitudinea existenta asupra unei valori punctuale a variabilei dependente

$\Rightarrow$ daca *intr-un* orar se cheltuie \$100000 pe publicitate TV si \$20000 pe publicitate radio, cu o incredere de 95%, valoarea vanzarilor va fi situata intre \$7929.6 si \$14583.3

Realizarea predictiilor (II)

- intervalul de incredere masoara incertitudinea existenta asupra valorii medii a variabilei dependente

$\Rightarrow$ daca *in fiecare* oras se cheltuie \$100000 pe publicitate TV si \$20000 pe publicitate radio, cu o incredere de 95%, valoarea (medie) a vanzarilor va fi situata intre \$10985.2 si \$11527.7

- intervalul de predictie masoara incertitudinea existenta asupra unei valori punctuale a variabilei dependente

$\Rightarrow$ daca *intr-un* orar se cheltuie \$100000 pe publicitate TV si \$20000 pe publicitate radio, cu o incredere de 95%, valoarea vanzarilor va fi situata intre \$7929.6 si \$14583.3

- intervalul de predictie *este mai larg* in comparatie cu cel de incredere, deoarece reflecta *incertitudinea (sporita)* a predictiei realizate pentru o instanta particulara in comparatie cu valoarea medie a variabilei dependente, pentru mai multe instante

Adaugarea unui termen de interactiune

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{TV} \times \texttt{radio}) \quad (7)$$

Adaugarea unui termen de interactiune

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{TV} \times \texttt{radio}) \tag{7}$$

Cod R

```
mod_interaction <- lm(data = Advertising, sales ~ TV * radio)
```

Adaugarea unui termen de interactiune

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{TV} \times \texttt{radio}) \quad (7)$$

Cod R

```
mod_interaction <- lm(data = Advertising, sales ~ TV * radio)
```

Rezultate

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV * radio	0.0011	0.000	20.73	< 0.0001

Interpretare

Sinergia dintre variabile

- β_3 ne arata cresterea eficacitatii cheltuielii cu publicitate TV, datorata cresterii cu o unitate a bugetului de publicitate radio
- o crestere cu \$1000 a publicitatii TV este asociata cu o crestere a vanzarilor cu $\hat{\beta}_1 + \hat{\beta}_3 \times radio = 19 + 1.1 \times radio$ unitati
- o crestere cu \$1000 a publicitatii radio este asociata cu o crestere a vanzarilor cu $\hat{\beta}_2 + \hat{\beta}_3 \times TV = 29 + 1.1 \times TV$ unitati

Compararea modelelor

Model	R^2	**RSE**	**F statistics**
Sales on TV and radio with interaction	0.9678	0.9435	1963
Sales on TV and radio	0.8972	1.681	859.6
Sales on TV	0.6119	3.259	312.1

Compararea modelelor

Model	R^2	**RSE**	**F statistics**
Sales on TV and radio with interaction	0.9678	0.9435	1963
Sales on TV and radio	0.8972	1.681	859.6
Sales on TV	0.6119	3.259	312.1

Modelul cel mai bun (dintre cele testate)

Este modelul prin care se explica vanzarile in functie de cheltuielile cu publicitatea TV si radio, considerandu-se si sinergia dintre acestea

Raspunsuri pentru intrebarile de cercetare

Raspunsuri pentru intrebarile de cercetare

Exista vreo relatie intre vanzari si bugetul de publicitate?

Da. Am rejectat ipoteza nula $H_0 : \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0$

Raspunsuri pentru intrebarile de cercetare

Exista vreo relatie intre vanzari si bugetul de publicitate?

Da. Am rejectat ipoteza nula $H_0 : \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0$

Cat de puternica este relatia?

R^2 este aproape 90%. RSE este 1.681, pentru o medie a vanzarilor de 14.022, care este aproximativ 12%

Raspunsuri pentru intrebarile de cercetare

Exista vreo relatie intre vanzari si bugetul de publicitate?

Da. Am rejectat ipoteza nula $H_0 : \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0$

Cat de puternica este relatia?

R^2 este aproape 90%. RSE este 1.681, pentru o medie a vanzarilor de 14.022, care este aproximativ 12%

Care dintre canalele de publicitate contribuie la vanzari?

Doar p-values la TV si radio sunt mici, deci doar publicitatea la TV si radio influenteaza vanzarile.

Raspunsuri pentru intrebarile de cercetare

Exista vreo relatie intre vanzari si bugetul de publicitate?

Da. Am rejectat ipoteza nula $H_0 : \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0$

Cat de puternica este relatia?

R^2 este aproape 90%. RSE este 1.681, pentru o medie a vanzarilor de 14.022, care este aproximativ 12%

Care dintre canalele de publicitate contribuie la vanzari?

Doar p-values la TV si radio sunt mici, deci doar publicitatea la TV si radio influenteaza vanzarile.

Cat de mare este efectul fiecarui canal de publicitate pentru vanzari?

Exista o asociere puternica intre publicitatea TV si vanzari si intre publicitatea radio si vanzari. Cand publicitatea la TV si radio sunt ignorate, atunci exista o asociere slaba intre publicitatea in presa scrisa si vanzari.

Raspunsuri pentru intrebarile de cercetare

Cu ce acuratete putem prezice vanzarile viitoare?

Daca dorim sa prezicem o valoare medie, putem calcula intervalul de incredere. Daca dorim sa prezicem un anume raspuns particular, vom calcula intervalul de predictie, care este mai larg

Raspunsuri pentru intrebarile de cercetare

Cu ce acuratete putem prezice vanzarile viitoare?

Daca dorim sa prezicem o valoare medie, putem calcula intervalul de incredere. Daca dorim sa prezicem un anume raspuns particular, vom calcula intervalul de predictie, care este mai larg

Este relatia determinata liniara?

Am observat un efect non-liniar de interactiune intre predictori care imbunatateste modelul de regresie

Raspunsuri pentru intrebarile de cercetare

Cu ce acuratete putem prezice vanzarile viitoare?

Daca dorim sa prezicem o valoare medie, putem calcula intervalul de incredere. Daca dorim sa prezicem un anume raspuns particular, vom calcula intervalul de predictie, care este mai larg

Este relatia determinata liniara?

Am observat un efect non-liniar de interactiune intre predictori care imbunatateste modelul de regresie

Exista o sinergie intre canalele de publicitate?

Valoarea p-value pentru termenul de interactiune indica prezenta unei asemenea sinergii. R^2 creste subtantial prin adaugarea termenului de interactiune