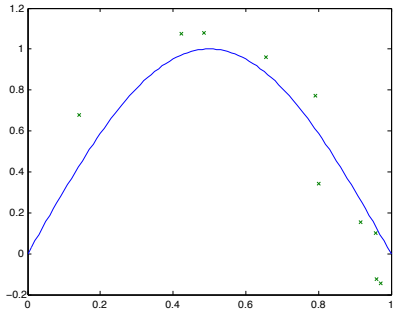
Problème de régression

On observe des données

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N) \in \mathbb{R}^n \times \mathbb{R}$$

On cherche à exprimer la dépendance entre $\mathbf{x}$ et y par une fonction.



Un exemple : USCrime data (L. Wasserman)

1. R: Crime rate: # of offenses reported to police per million population
2. Age: The number of males of age 14-24 per 1000 population
3. S: Indicator variable for Southern states (0 = No, 1 = Yes) Ed: Mean # of years of schooling x 10 for persons of age 25 or older
4. Ex0: 1960 per capita expenditure on police by state and local government
5. Ex1: 1959 per capita expenditure on police by state and local government
6. LF: Labor force participation rate per 1000 civilian urban males age 14-24
7. M: The number of males per 1000 females
8. N: State population size in hundred thousands
9. NW: The number of non-whites per 1000 population
10. U1: Unemployment rate of urban males per 1000 of age 14-24
11. U2: Unemployment rate of urban males per 1000 of age 35-39
12. W: Median value of transferable goods and assets or family income in tens of $
13. X: The number of families per 1000 earning below 1/2 the median income

Collected from the FBI's Uniform Crime Report and other government agencies to determine how the variable crime rate depends on the other variables measured in the study.

Un exemple : USCrime (suite)

R	Age	S	Ed	Ex0	Ex1	LF	M	N	NW	U1	U2	W	X
79.1	151	1	91	58	56	510	950	33	301	108	41	394	261
163.5	143	0	113	103	95	583	1012	13	102	96	36	557	194
57.8	142	1	89	45	44	533	969	18	219	94	33	318	250
196.9	136	0	121	149	141	577	994	157	80	102	39	673	167
123.4	141	0	121	109	101	591	985	18	30	91	20	578	174
68.2	121	0	110	118	115	547	964	25	44	84	29	689	126
96.3	127	1	111	82	79	519	982	4	139	97	38	620	168
155.5	131	1	109	115	109	542	969	50	179	79	35	472	206
85.6	157	1	90	65	62	553	955	39	286	81	28	421	239
70.5	140	0	118	71	68	632	1029	7	15	100	24	526	174
167.4	124	0	105	121	116	580	966	101	106	77	35	657	170
84.9	134	0	108	75	71	595	972	47	59	83	31	580	172
51.1	128	0	113	67	60	624	972	28	10	77	25	507	206
66.4	135	0	117	62	61	595	986	22	46	77	27	529	190
79.8	152	1	87	57	53	530	986	30	72	92	43	405	264
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Expliquer la variable R par les autres attributs.

Modélisation de la régression

- Une variable aléatoire $Z = (X, Y)$ à valeurs dans $\mathbb{R}^n \times \mathbb{R}$
- Les exemples sont des couples $(x, y) \in \mathbb{R}^n \times \mathbb{R}$ tirés selon la distribution jointe

$$P(Z = (x, y)) = P(X = x)P(Y = y|X = x).$$

- Un échantillon S est un ensemble fini d'exemples

$$\{(x_1, y_1), \ldots, (x_N, y_N)\}$$

i.i.d. selon P.

Modélisation de la régression (suite)

On cherche une fonction : $f : \mathbb{R}^n \to \mathbb{R}$.

Fonction de perte (loss function)

$$L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2.$$

La fonction risque (ou erreur) : espérance mathématique de la fonction de perte.

$$R(f) = \int L(y, f(\mathbf{x}))dP(\mathbf{x}, y) = \int_{\mathbb{R}^n \times \mathbb{R}} (y - f(\mathbf{x}))^2 dP(\mathbf{x}, y).$$

Le problème général de la régression :

étant donné un échantillon $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, trouver une fonction f de risque $R(f)$ minimal.

La fonction de régression

Définition : *fonction de régression*

$$r(\mathbf{x}) = \int_Y y dP(y|\mathbf{x}) = \mathbb{E}_y(y|\mathbf{x})$$

Pour chaque $\mathbf{x}$, $r(\mathbf{x})$ est égal à la moyenne des observations

Théorème : la fonction de régression minimise le risque quadratique.

$$r = Argmin_f R(f)$$

Remarque : Comme la fonction de Bayes en classification, la fonction de régression est le plus souvent inaccessible. On cherche à l'approcher.

Minimisation du risque empirique

Soit $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ un ensemble d'observations.

- On cherche à approcher la fonction qui minimise $R(f)$
- Le risque empirique $R_{emp}(f)$ de f est la moyenne des carrés des écarts à la moyenne de f calculée sur S :

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2.$$

- Principe de minimisation du risque empirique : calculer

$$\underset{f}{ArgMin}\ R_{emp}(f)$$

Méthode des moindres carrés.

Régression linéaire simple

On suppose que

- X prend des valeurs dans $\mathbb{R}$: cas où $n = 1$,
- $Y = w_0 + w_1 X + \epsilon$ où ϵ est un bruit aléatoire vérifiant
  - $\mathbb{E}(\epsilon) = 0$ et
  - $\mathbb{V}(\epsilon) = \sigma^2$ (variance indépendante de X).

La fonction de régression est

$$r(x) = w_0 + w_1 x.$$

Régression linéaire simple : estimateurs des moindres carrés

Soit $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ un échantillon, où chaque $(x_i, y_i) \in \mathbb{R}^2$.
Soit $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = w_0 + w_1 x$.

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + w_1 x_i))^2$$

Théorème. Les valeurs de $\widehat{w_1}$ et $\widehat{w_0}$ qui minimisent $R_{emp}(f)$ sont

$$\widehat{w_1} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \text{ et } \widehat{w_0} = \overline{y} - \widehat{w_1}\overline{x}$$

où

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \text{ et } \overline{y} = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

Régression linéaire : estimateurs des moindres carrés (suite)

La fonction de régression estimée est alors

$$\widehat{r}(x) = \widehat{w_1} x + \widehat{w_0}.$$

Les erreurs estimées sont

$$\widehat{\epsilon_i} = y_i - \widehat{y_i} = y_i - (\widehat{w_1} x_i + \widehat{w_0}).$$

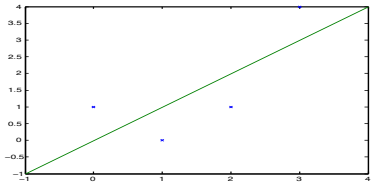
La variance estimée est

$$\widehat{\sigma}^2 = \frac{1}{N-2} \sum_{i=1}^{N} \widehat{\epsilon_i}^2.$$

Estimateurs des moindres carrés : exemple

Soit $S = \{(0, 1), (1, 0), (2, 1), (3, 4))\}$ un échantillon.
On trouve

$$\overline{x} = 3/2, \overline{y} = 3/2, \widehat{w_1} = 1 \text{ et } \widehat{w_0} = 0.$$



On a $\widehat{\epsilon_1} = 1, \widehat{\epsilon_2} = -1, \widehat{\epsilon_3} = -1, \widehat{\epsilon_4} = 1$ et $\widehat{\sigma}^2 = 2$.

Propriétés de l'estimateur des moindres carrés

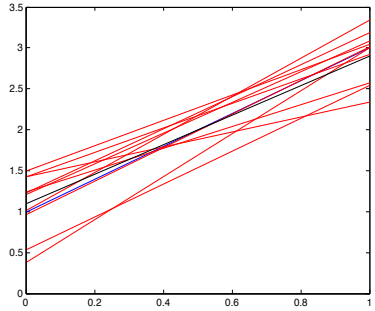
- $\widehat{w_1}, \widehat{w_0}$ et $\widehat{\sigma}^2$ sont des *estimateurs non biaisés* de w_1, w_0 et σ^2 : si l'on répète un grand nombre d'expériences avec le même modèle, les moyennes des estimations convergent vers les paramètres du modèle.
- $\widehat{w_1}, \widehat{w_0}$ et $\widehat{\sigma}^2$ sont des *estimateurs consistants* de w_1, w_0 et σ^2 : plus on dispose d'observations, plus les estimations se rapprochent des paramètres du modèle.
- si ϵ suit une loi normale, l'estimateur des moindres carrés est aussi l'*estimateur du maximum de vraisemblance* : celui qui maximise la probabilité des observations.

Estimateur non biaisé : illustration

X prend 11 valeurs équidistantes dans $[0, 1]$; $Y = 2 * X + 1 + Norm(0, 1)$.

On réalise 10 expériences.

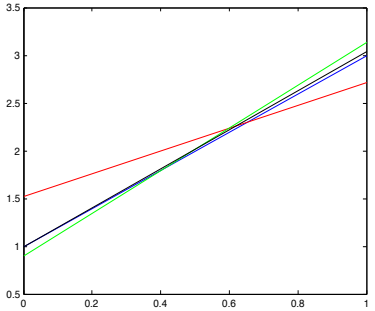
- en bleu : la droite de régression
- en rouge : chaque estimation
- en noir : la moyenne des estimations.



Estimateur consistant : illustration

X prend N valeurs équidistantes dans $[0, 1]$; $Y = 2 * X + 1 + Norm(0, 1)$.

- en bleu : la droite de régression
- en rouge : $N = 11$
- en vert : $N = 101$
- en noir : $N = 1001$.



Moindre carrés et maximum de vraisemblance

Supposons que ϵ suive une loi Normale $\mathcal{N}(0, \sigma^2)$ de moyenne 0 et de variance σ^2, de densité

$$f(x) = \frac{1}{\sigma\sqrt{\pi}} exp\left\{-\frac{x^2}{2\sigma^2}\right\}$$

Pour une fonction $x \mapsto w_0 + w_1 x$, la vraisemblance de l'échantillon $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ est égale à

$$L_S(w_0, w_1) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{\pi}} exp\left\{-\frac{(y_i - w_0 - w_1 x_i)^2}{2\sigma^2}\right\}$$

Théorème : l'estimateur des moindres carrés est aussi la fonction qui maximise la vraisemblance de l'observation.

Régression linéaire multiple (version 1)

On suppose que

- X prend ses valeurs dans $\mathbb{R}^n$,
- $Y = w_0 + \mathbf{w}^\top X + \epsilon$ où
  - $w_0 \in \mathbb{R}$ et $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$,
  - ϵ est une variable aléatoire telle que
    - $\mathbb{E}(\epsilon) = 0$ et
    - $\mathbb{V}(\epsilon) = \sigma^2$ (variance indépendante de X).

Remarque : pour homogénéiser les notations, on supposera que la première coordonnée de X est toujours égale à 1 ; cela permet, en rentrant w_0 dans le vecteur $\mathbf{w}$, d'avoir un modèle plus simple et aussi expressif.

Régression linéaire multiple (version 2)

On suppose que

- X prend ses valeurs dans $\{1\} \times \mathbb{R}^n$
- $Y = \mathbf{w}^\top X + \epsilon$ où
  - $\mathbf{w} = (w_0, w_1, \ldots, w_n) \in \mathbb{R}^{n+1}$,
  - ϵ est une variable aléatoire telle que
    - $\mathbb{E}(\epsilon) = 0$ et
    - $\mathbb{V}(\epsilon) = \sigma^2$ (variance indépendante de X).

La fonction de régression est

$$r(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} = w_0 + w_1 x_1 + \ldots w_n x_n.$$

Régression linéaire multiple : estimateur des moindres carrés

Soit $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\} \subset \mathbb{R}^{n+1}$ l'échantillon d'apprentissage.

Soit X la matrice $I \times (n+1)$ dont la i-ème ligne est : $1, \boldsymbol{x}_i$.
Soit Y le vecteur colonne composé des étiquettes y_i.

L'estimateur des moindres carrés est

$$\mathbf{w} = (X^\top X)^{-1} X^\top Y$$

où $X^\top$ désigne la matrice transposée de X.

Si $X^\top X$ n'est pas inversible, ou si $det(X^\top X) \simeq 0, \ldots$ il est nécessaire de transformer le problème.

Estimateurs des moindres carrés : exemple

Exemple : $S = \{((0,0),-1), ((0,1),1), ((1,0),1), ((1,1),1)\}$.
On a

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}, X^\top = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix} \text{ et } Y = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

On vérifie que

$$X^\top X = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{pmatrix}, (X^\top X)^{-1} = \begin{pmatrix} 3/4 & -1/2 & -1/2 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{pmatrix}$$

$$(X^\top X)^{-1} X^\top = \begin{pmatrix} 3/4 & 1/4 & 1/4 & -1/4 \\ -1/2 & -1/2 & 1/2 & 1/2 \\ -1/2 & 1/2 & -1/2 & 1/2 \end{pmatrix} \text{ et } (X^\top X)^{-1} X^\top Y = \begin{pmatrix} -1/2 \\ 1 \\ 1 \end{pmatrix}$$

soit

$$w_0 = -1/2, w_1 = w_2 = 1.$$

Modèles linéaires généralisés

- Peu vraisemblable, en général, que les données d'observations $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{n+1}$ puissent être précisément décrites par un modèle linéaire.
- Sous des conditions de régularités très générales, toute fonction f peut être approchée, au moins localement, par une combinaison **linéaire**
  - de *monômes* (développements limités)

$$\sin x = x - x^3/3 + x^5/5 - \ldots$$

  - ou de *fonctions trigonométriques* (développements de Fourier)

$$x = 2\left(\sin x - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \ldots\right)$$

Modèles linéaires généralisés (suite)

Soient $\phi_1, \ldots, \phi_M : \mathbb{R}^n \mapsto \mathbb{R}$ un ensemble de *fonctions de base*.

Exemples :

- Fonctions coordonnées : $M = n$ et $\phi_i(\mathbf{x}) = x_i$.
- Fonctions polynomes : $n = 1$ et $\phi_i(x) = x^i$.
- Fonctions Gaussiennes : $n = 1$ et $\phi_i(x) = exp(-(x - \mu_i)^2/\sigma^2)$.

On transforme

- chaque $\boldsymbol{x}$ en un vecteur $\boldsymbol{\phi}(\mathbf{x}) = (\phi_0(\boldsymbol{x}), \ldots, \phi_M(\boldsymbol{x})) \in \mathbb{R}^M$
- l'échantillon $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ en $(\boldsymbol{\phi}(\mathbf{x}_1), y_1), \ldots, (\boldsymbol{\phi}(\mathbf{x}_N), y_N)$

Remarques :

- On suppose souvent que les fonctions de base contiennent une fonction constante $\phi_0(\mathbf{x}) = 1$.
- Dans le cas des fonctions coordonnées, on retrouve le modèle de la régression multiple.

Modèles linéaires généralisés (suite)

Échantillon : $(\boldsymbol{\phi}(\mathbf{x}_1), y_1), \ldots, (\boldsymbol{\phi}(\mathbf{x}_N), y_N)$

On considère le problème suivant : trouver la fonction

$$f(\mathbf{x}) = \alpha_0\phi_0(\mathbf{x}) + \cdots + \alpha_M\phi_M(\mathbf{x})$$

de risque empirique

$$R_{emp}(f) = \sum_{i=1}^{N}\left(y_i - \left(\sum_{j=0}^{M}\alpha_j\phi_j(\boldsymbol{x}_i)\right)\right)^2 = \sum_{i=1}^{N}(y_i - \boldsymbol{\alpha}^\top\boldsymbol{\phi}(\mathbf{x}_i)^2)$$

minimal.

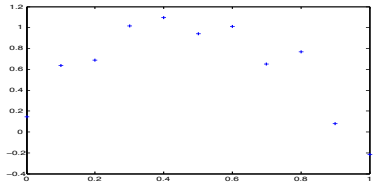
Par régression multiple, on obtient une fonction f qui est

- non linéaire par rapport à $\mathbf{x}$,
- linéaire par rapport aux paramètres $\boldsymbol{\alpha}$ du modèle.

Un exemple

On observe les données suivantes :

0	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000	1.0000
0.1434	0.6351	0.6856	1.0160	1.0964	0.9393	1.0098	0.6516	0.7655	0.0796	-0.2138



Les données ne semblent pas alignées : on les transforme par les fonctions

$$\phi_0(x) = 1, \phi_1(x) = x, \phi_2(x) = x^2 \text{ et } \phi_3(x) = x^3.$$

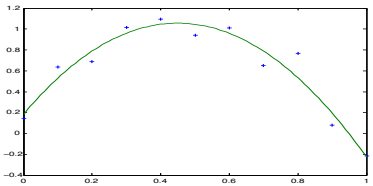
Un exemple (suite)

Une régression linéaire généralisée permet de trouver

$$\alpha_0 = 0.1848, \alpha_1 = 3.8960, \alpha_2 = -4.3942 \text{ et } \alpha_3 = 0.0878$$

soit le polynôme

$$p(x) = 0.1848 + 3.8960x - 4.3942x^2 + 0.0878x^3.$$



Sélection de variables

Les observations x peuvent dépendre d'un très grand nombre de variables .

- Constat : trop de variables nuisent à la qualité de la prédiction.
- Le modèle induit peut être difficile à interpréter s'il fait intervenir toutes les variables au même niveau.
- Question : comment trouver un bon compromis entre biais et variance ? *underfitting* et *overfitting* ?

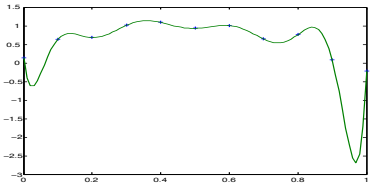
En trouvant un compromis entre

- l'adéquation aux données et
- la complexité du modèle.

Un exemple (suite)

Si l'on prend comme base les monômes $\phi_i(x) = x^i$ pour $0 \leq i \leq 10$, on trouve le polynôme

$$p(x) \simeq 0.1463 - 76.0471x + 2394.5864x^2 - 28139.8810x^3 + 173816.7283x^4 - 634281.2333x^5 + 1437424.4162x^6 - 2044405.0970x^7 + 1774215.9551x^8 - 858102.0064x^9 + 177152.2278x^{10}.$$



Quels sont les monômes les plus significatifs ?

Sélection de variables

- Soit $\mathcal{V}$ l'ensemble de variables et soit $W \subseteq \mathcal{V}$ contenant d variables.
- Soit $\psi_W : \mathbb{R}^n \mapsto R^d$ la projection sur les coordonnées de W.
- On transforme l'échantillon en $S_W = \{(\psi_W(\mathbf{x}_1)), y_1), \ldots, (\psi_W(\mathbf{x}_N)), y_N)\}$.
- Soit $\widehat{r}_W$ la fonction de risque empirique minimal calculée sur S_W.
- On note l_W la log-vraisemblance du modèle $\widehat{r}_W$:

$$l_W \simeq -\sum_{i=1}^{N} (y_i - \widehat{r}_W(\psi_W(\mathbf{x}_i))^2$$

On cherche par exemple l'ensemble W pour lequel

$$l_W - d$$

est maximal (méthode AIC, pour Akaike Information Criterion).

- l_W est un terme qui mesure l'adéquation aux données
- d est une mesure de complexité du modèle.

Trouver un compromis entre l'adéquation aux données et la complexité du modèle.

Deux stratégies pour sélectionner W

Pour sélectionner W, on peut envisager tous les ensembles W possibles et évaluer W selon le critère précédent. Mais c'est impraticable si le nombre de variables est trop important !

Deux stratégies alternatives gloutonnes : forward and backward stepwise regression.

- Partir du modèle vide. Ajouter au modèle courant la variable qui fait croître le plus le score associé au critère choisi. Arrêter lorsqu'aucune variable ne fait plus croître le score.
- De manière symétrique, partir de l'ensemble des variables $\mathcal{V}$ et les supprimer l'une après l'autre, tant que le score associé au critère choisi croît.

Ridge regression

Autre idée : pénaliser la taille des paramètres. On cherche à minimiser

$$\sum_{i=1}^{N}(y_i - \boldsymbol{\alpha}^\top \mathbf{x}_i)^2 + C||\boldsymbol{\alpha}||^2.$$

- Se résout aussi simplement que la régression multivariée sans pénalisation :

$$\boldsymbol{\alpha} = (\mathbf{X}^\top\mathbf{X} + C\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

  où $\mathbf{I}$ est la matrice identité de dimension n.
- Un des avantages de la régression ridge est que la matrice $(\mathbf{X}^\top\mathbf{X} + C\mathbf{I})$ est toujours non singulière.
- Mais la solution n'est plus invariante par homothétie ou par translation. En pratique, on *normalise* (ou on *standardise*) souvent les données de façon qu'elles aient une moyenne nulle (et chaque attribut une variance égale à 1).
- Mais comment trouver la valeur de C ? Par exemple, par validation croisée.

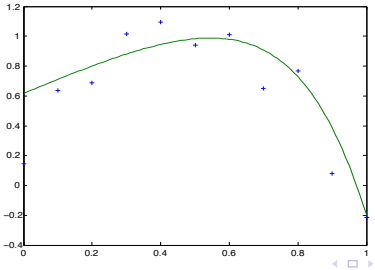
Un exemple (suite)

Avec les monômes $\phi_i(x) = x^i$ pour $0 \leq i \leq 10$ et $C = 0.1$, on trouve les coefficients

$$0.6189\ ;\ 0.9368\ ;\ -0.0655\ ;\ -0.3620\ ;\ -0.4026$$
$$-0.3529\ ;\ -0.2745\ ;\ -0.1906\ ;\ -0.1095\ ;\ -0.0341\ ;\ 0.0350$$

et donc le polynome

$$p(x) \simeq 0.6189 - 0.9368x - 0.0655x^2 - 0.3620x^3 + -0.4026x^4 - 0.3529x^5$$
$$-0.2745x^6 - 0.1906x^7 + -0.1095x^8 - 0.0341x^9 + 0.0350x^{10}.$$



Lasso

Autre idée : on cherche à minimiser

$$\sum_{i=1}^{N}(y_i - \boldsymbol{\alpha}^\top \mathbf{x}_i)^2 + C||\boldsymbol{\alpha}||_1$$

où

$$||\boldsymbol{\alpha}||_1 = |\alpha_0| + \cdots + |\alpha_M|.$$

- Plus difficile à résoudre. Par exemple, par *programmation quadratique*.
- Un grand nombre de coefficients s'annulent : modèle parcimonieux.
- Il faut toujours déterminer une bonne valeur pour C - par exemple, par validation croisée.

Un exemple (suite)

Avec les monômes $\phi_i(x) = x^i$ pour $0 \leq i \leq 10$ et $C = 0.01$, on trouve les coefficients

$$0.55746891\ ;\ 0.68424671\ ;\ 0\ ;\ 0\ ;\ 0\ ;\ -1.44786439\ ;\ 0\ ;\ 0\ ;\ 0\ ;\ 0\ ;\ 0$$

et donc le polynome

$$p(x) \simeq 0.55746891 + 0.68424671x - 1.44786439x^5$$