Cours MPC - Partie 2

Régression : prédiction d'une variable à l'aide d'une ou plusieurs autres variables

Simon Malinowski

M1 Miage, Univ. Rennes 1

Des données aux prédictions

Exemple de régression

Maison	Taille de la maison (pieds carrés)	Année de construction	Prix de l'immobilier (objectif, en milliers)e)
1	80	1985	120
2	92	2010	180
3	75	2008	145
4	110	2015	220
5	85	2000	140
6	150	1975	225
7	55	1992	100
8	105	1999	? ?
9	95	2018	? ?

Autres exemples de régression

- Combien de personnes retweeteront votre tweet ? (y)
  - → dépend de x : nombre de followers, popularité, sujet du tweet, ...
- Quel sera votre prochain salaire ? (y)
  - → dépend de x : votre diplôme, votre expérience, ...
- Combien de points aura une équipe de football à la fin de la saison ?
  - → dépend de x : statistiques sur les performances de l'équipe (buts, tirs, possession,...)

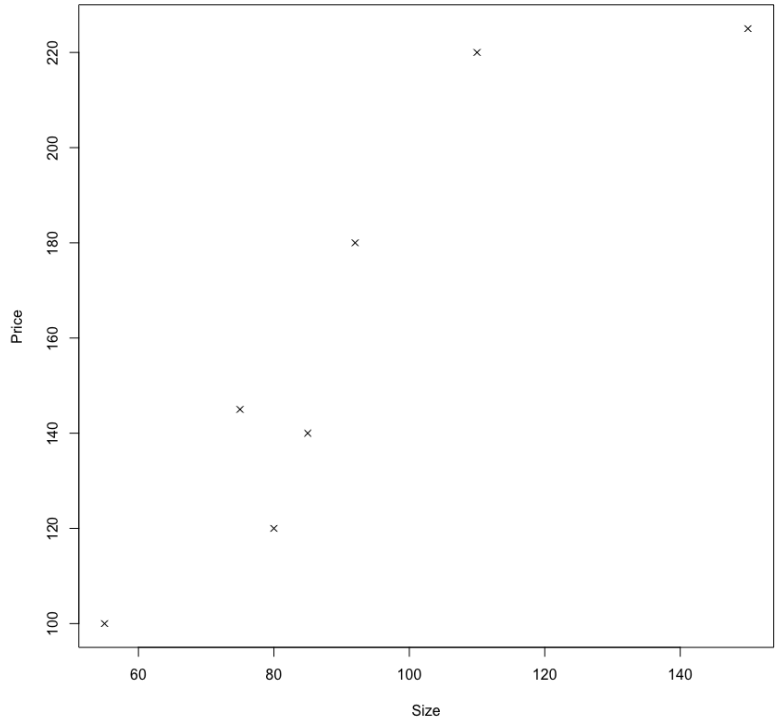
Différents types de régression

Régression →la variable cible est numérique

- Simple VS multiple
  - → régression simple : une variable caractéristique (X) pour prédire Y
  - → régression multiple : plusieurs variables caractéristiques ($X_1, \ldots, X_p$) pour prédire Y
- Linéaire VS Non linéaire

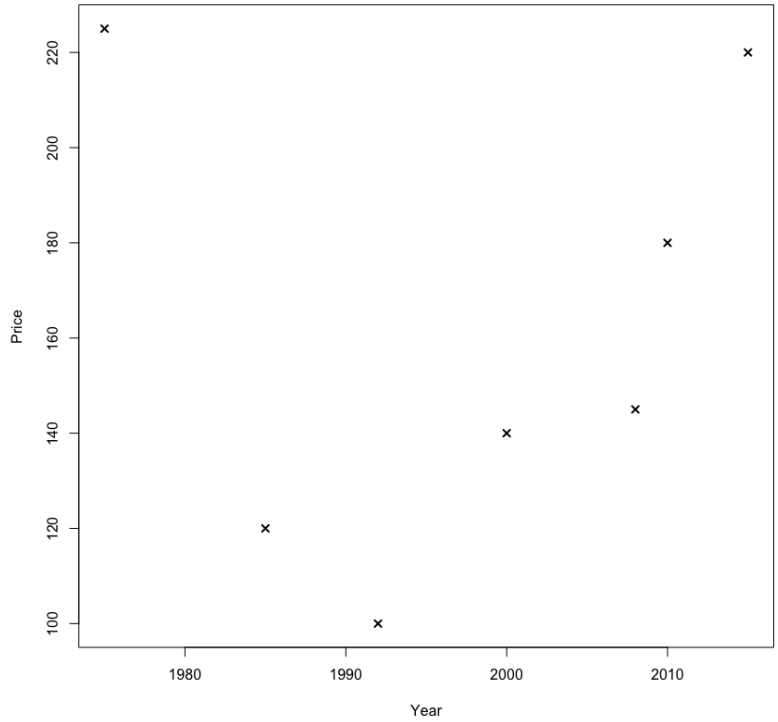
Régression linéaire simple en quelques mots

Taille	Prix
80	120
92	180
75	145
110	220
85	140
150	225
55	100
105	??
95	??



Régression linéaire simple en quelques mots

Année	Prix
1985	120
2010	180
2008	145
2015	220
2000	140
1975	225
1992	100
1999	??
2018	??



Ce que nous verrons sur la régression simple

- Comment trouver le modèle ? (ie coefficients de la droite de régression)
- Comment évaluer la performance d'un modèle ?
  - → le modèle sera-t-il bon pour prédire de nouvelles entrées ?
  - → si j'ai plus d'une variable, laquelle me semble la meilleure ?

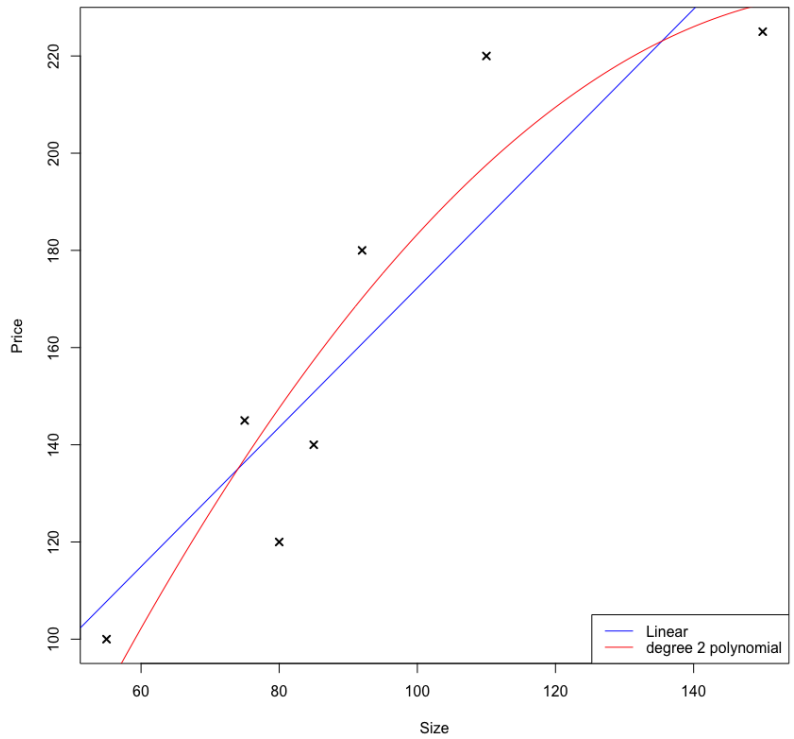
La régression linéaire multiple en bref

Taille	Année	Prix
80	1985	120
92	2010	180
75	2008	145
110	2015	220
85	2000	140
150	1975	225
55	1992	100
105	1999	??
95	2018	??

$$P = -2872 + 1.65 \times S + 1.44 \times Oui$$

Régression linéaire multiple pour la régression non linéaire

Taille	$Taille^2$	Prix
80	6400	120
92	8464	180
75	5625	145
110	12100	220
85	7225	140
150	22500	225
55	3025	100
105	11025	??
95	9025	??



$$P = -91.3 + 3.94 \times S - 0.012 \times S^2$$

Ce que nous verrons à propos de la régression multiple

- Comment trouver le modèle ? (ie les coefficients du modèle)
- Comment évaluer la performance d'un modèle ?
  - → le modèle sera-t-il bon pour prédire de nouvelles entrées ?
  - → comment comparer des modèles avec un nombre différent de variables
- Régression non linéaire avec régression multiple
- Sélection de variables
  - → Ai-je vraiment besoin de toutes les variables dont je dispose, ou un sous-ensemble serait peut-être mieux ?
  - → comment effectuer une sélection de variables ? (critères, méthodes)

Régression avec des variables non numériques

Taille	District	Prix
80	Centre	120
92	Bourg-Lesvèque	180
75	Centre	145
110	Longchamps	220
85	Longchamps	140
150	Bourg-Lesvèque	225
55	Centre	100
105	Longchamps	? ?
95	Centre	? ?

Données, problème

Régression simple : 1 variable cible numérique, 1 variable explicative numérique

Individus	X	Oui
1	x_1	et_1
	...	
je	x_{je}	et_{je}
	...	
n	x_n	et_n

Régression multiple : 1 variable cible numérique, plusieurs variables explicatives numériques

Individus	X_1	...	X_j	...	X_p	*Oui*
1	x_{11}	...	x_{1j}	...	x_{1p}	et_1
	...		...		...	
je	x_{je1}	...	x_{ij}	...	*Xadresse IP*	et_{je}
	...		...		...	
n	x_{n1}	...	*XNew Jersey*	...	x_{np}	et_n

Quantité d'informations dans la variable cible

$$je_t = \sum_{je=1}^{n} (et_{je} - \bar{et})^2$$

Modèle linéaire

Nous supposons que Oui_i est une fonction linéaire des variables explicatives X_{je}

$$Y' \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Estimation des coefficients

$\beta_0, \ldots, \beta_p$ sont les paramètres « optimaux » (que nous trouverions si nous connaissions toutes les données sur le problème)

Objectif : trouver $\hat{\beta_0}, \ldots, \hat{\beta_p}$, estimations de $\beta_0, \ldots \beta_p$ qui sont aussi précis que possible

$\rightarrow$ Régression des moindres carrés

Régression des moindres carrés (cas de régression simple)

→ Objectif : minimiser la somme des carrés des résidus

$$je_r(\beta_0, \beta_1) = \sum_{je=1}^{n} (\hat{y}_{je} - et_{je})$$

$$je_r(\beta_0, \beta_1) = \sum_{je=1}^{n} ((\beta_0 + \beta_1 x_{je}) - et_{je})^2$$

Les dérivées partielles doivent être égales à 0 :

$$\begin{cases} \beta_1 = \frac{\sum (x_{je} - \bar{x})(et_{je} - \bar{y})}{(x_{je} - \bar{x})^2} \\ \beta_0 = \bar{y} - \beta_1 \bar{x} \end{cases}$$

Nous pouvons également calculer $\sigma_{\hat{\beta_1}}$:l'incertitude quant à la valeur de β_1

Coefficient de détermination

Nous avons vu que la quantité d'information dans la variable cible est :

$$je_t = \sum_{je=1}^{n} (et_{je} - \bar{et})^2$$

Cette quantité d'informations peut être divisée en deux parties :

- la quantité d'informations expliquées par le modèle je_m
- la quantité d'informations non expliquées par le modèle je_r

Nous avons $je_t = je_m + je_r$

Le R-carré est défini comme :

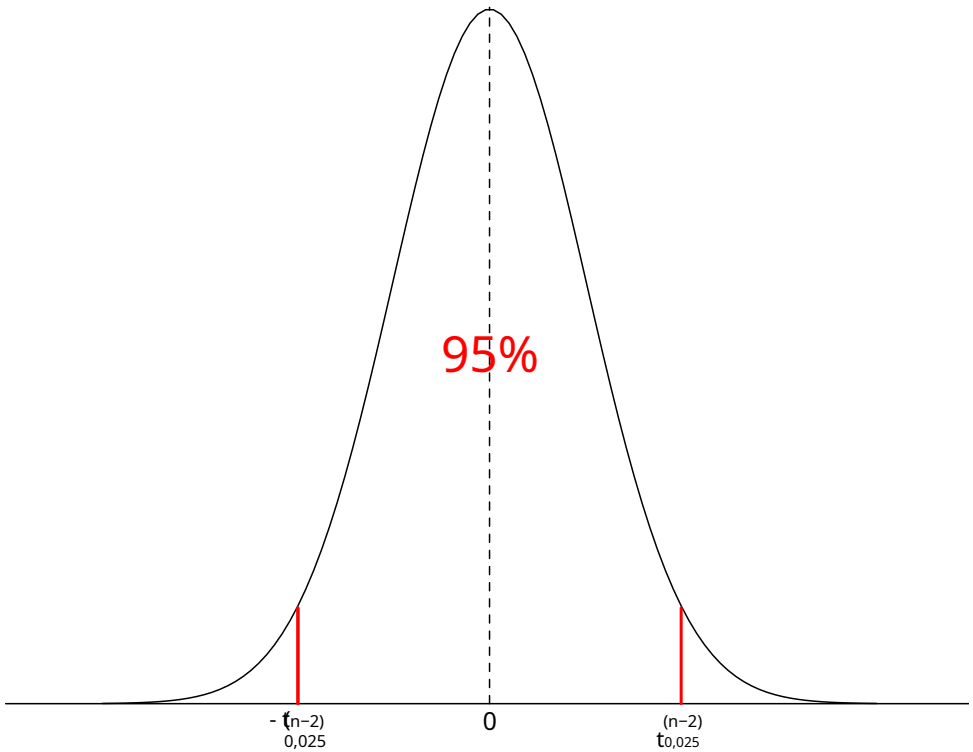
$$R^2 = m\frac{je}{je_t}$$

Régression simple : test de signification de β_1

Hypothèse 0 (H_0) : X n'influence pas Y $\rightarrow \beta_1 = 0$

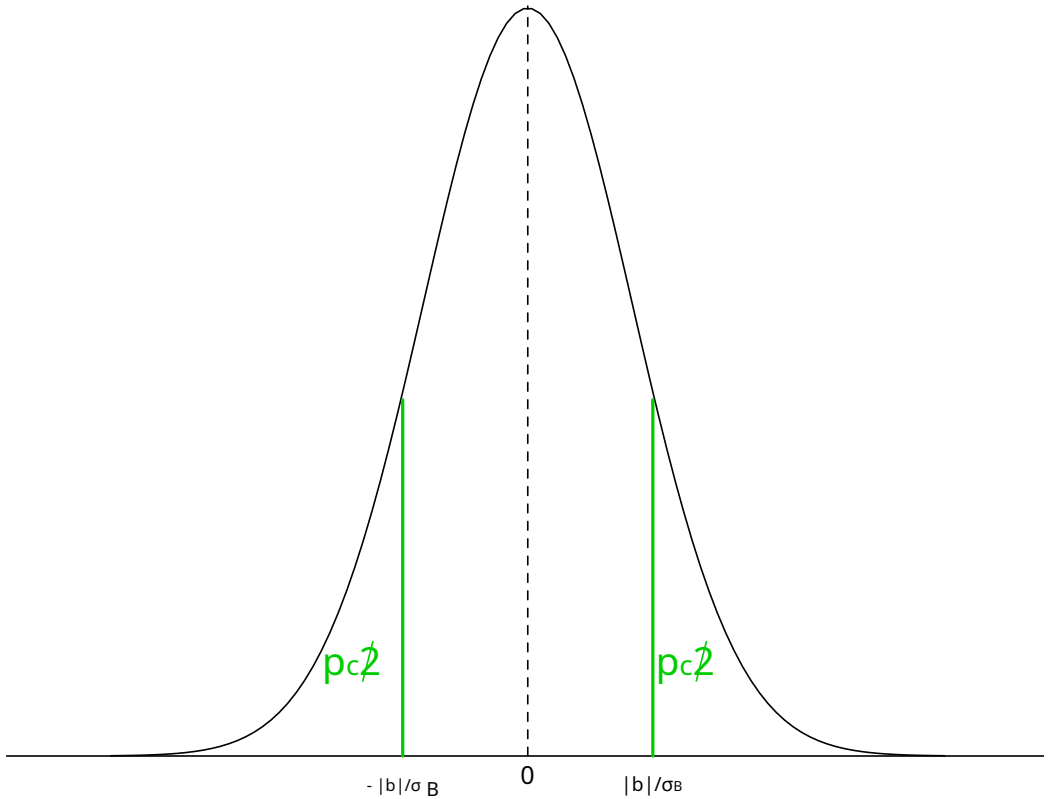
Hypothèse 1 (H_1) : X a une influence sur Y $\rightarrow \beta_1 \neq 0$

Sous H_0, $\frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_1}}$ est généré par une loi de Student avec $n-2$ degrés de liberté



Régression simple : test de signification de β_1

Sous H_0, $\frac{\hat{\beta_1}}{\sigma_{\hat{\beta_1}}}$ est généré par une loi de Student avec $n-2$ degrés de liberté



p_c : probabilité de prendre une mauvaise décision lors de la prise de décision H_1

$p_c < 0.05 \Rightarrow$ décider H_1

Multiple reg. : Test de significativité d'un coefficient

Hypothèse H_0 : $\beta_j = 0$ (x_j n'est pas significatif en présence d'autres variables)

Hypothèse H_1 : $\beta_j \neq 0$ (x_j est significatif en présence d'autres variables)

Valeur du test : $T = \frac{\hat{\beta}_j}{\hat{\sigma}_j}$

Sous H_0, T est généré par une loi de Student avec $(n - p - 1)$ degrés de liberté.

Si $|T| > t_{n-p-1}$, nous refusons H_0.

Exemples de l'intérêt de ce test

```
Call:
lm(formula = y ~ x1 + x10, data = ozo)

Residuals:
    Min      1Q  Median      3Q     Max
-41.940 -11.255   0.369   9.500  55.273

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.96058   10.15508  -2.852  0.00521 **
x1            4.47827    0.66801   6.704 9.58e-10 ***
x10           0.40446    0.07219   5.602 1.63e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.44 on 108 degrees of freedom
Multiple R-squared:  0.6275,    Adjusted R-squared:  0.6206
F-statistic: 90.98 on 2 and 108 DF,  p-value: < 2.2e-16
```

Exemples de l'intérêt de ce test

```
Call:
lm(formula = y ~ x1 + x2, data = ozo)

Residuals:
    Min      1Q  Median      3Q     Max
-38.822 -11.896   0.239  11.338  46.672

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -31.1244    10.2426  -3.039  0.00298 **
x1            0.8023     1.2252   0.655  0.51398
x2            4.9473     0.9254   5.346 5.06e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.61 on 108 degrees of freedom
Multiple R-squared:  0.6199,    Adjusted R-squared:  0.6128
F-statistic: 88.06 on 2 and 108 DF,  p-value: < 2.2e-16
```

Exemples de l'intérêt de ce test

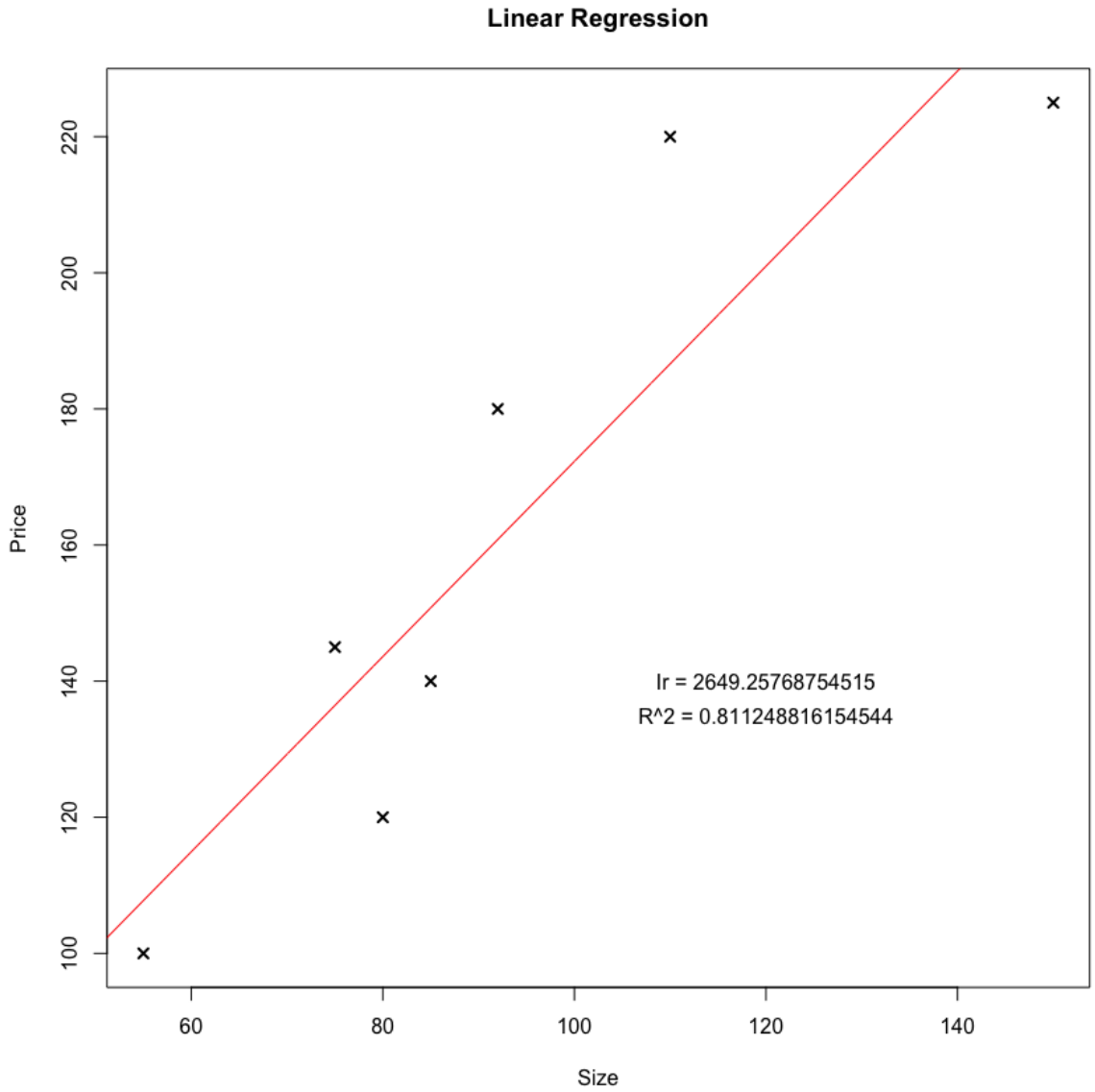
```
Call:
lm(formula = y ~ x1 + x10 + x2, data = ozo)

Residuals:
    Min      1Q  Median      3Q     Max
-49.982  -8.149   0.438   9.916  38.693

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.11029    9.04994  -3.106  0.00243 **
x1           -0.55031    1.10696  -0.497  0.62011
x10           0.36521    0.06474   5.641 1.39e-07 ***
x2            4.42587    0.82145   5.388 4.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.54 on 107 degrees of freedom
Multiple R-squared:  0.707, Adjusted R-squared:  0.6988
F-statistic: 86.07 on 3 and 107 DF,  p-value: < 2.2e-16
```

Un modèle donné est-il bon pour faire des prédictions ?

Sont le coefficient de détermination ou le *je ne sais pas* (ou les tests statistiques) adaptés pour me dire si un modèle donné est bon pour faire des prédictions ?

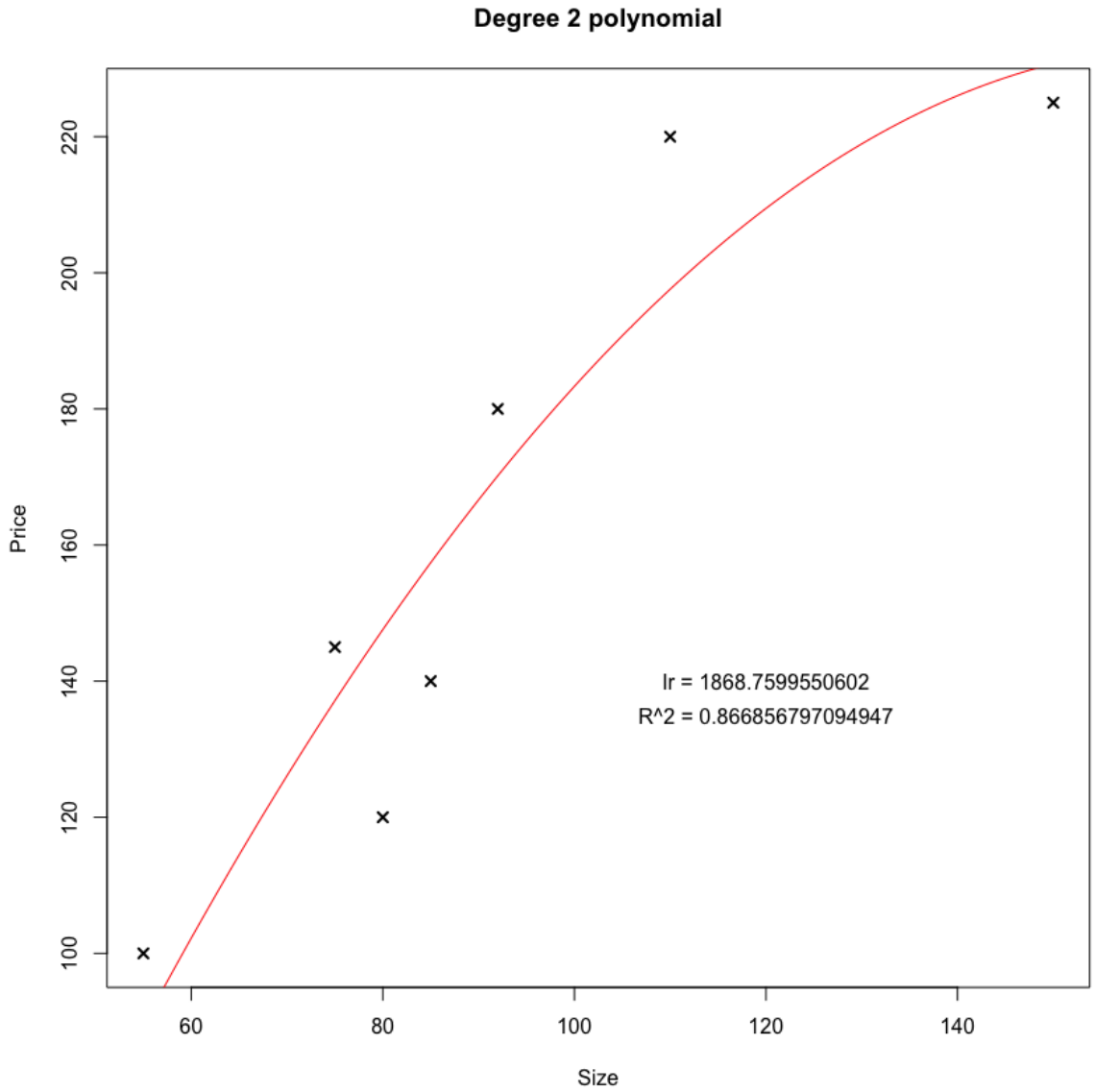
je_r et R^2 ne sont pas de bons indicateurs

Taille	Prix
80	120
92	180
75	145
110	220
85	140
150	225
55	100



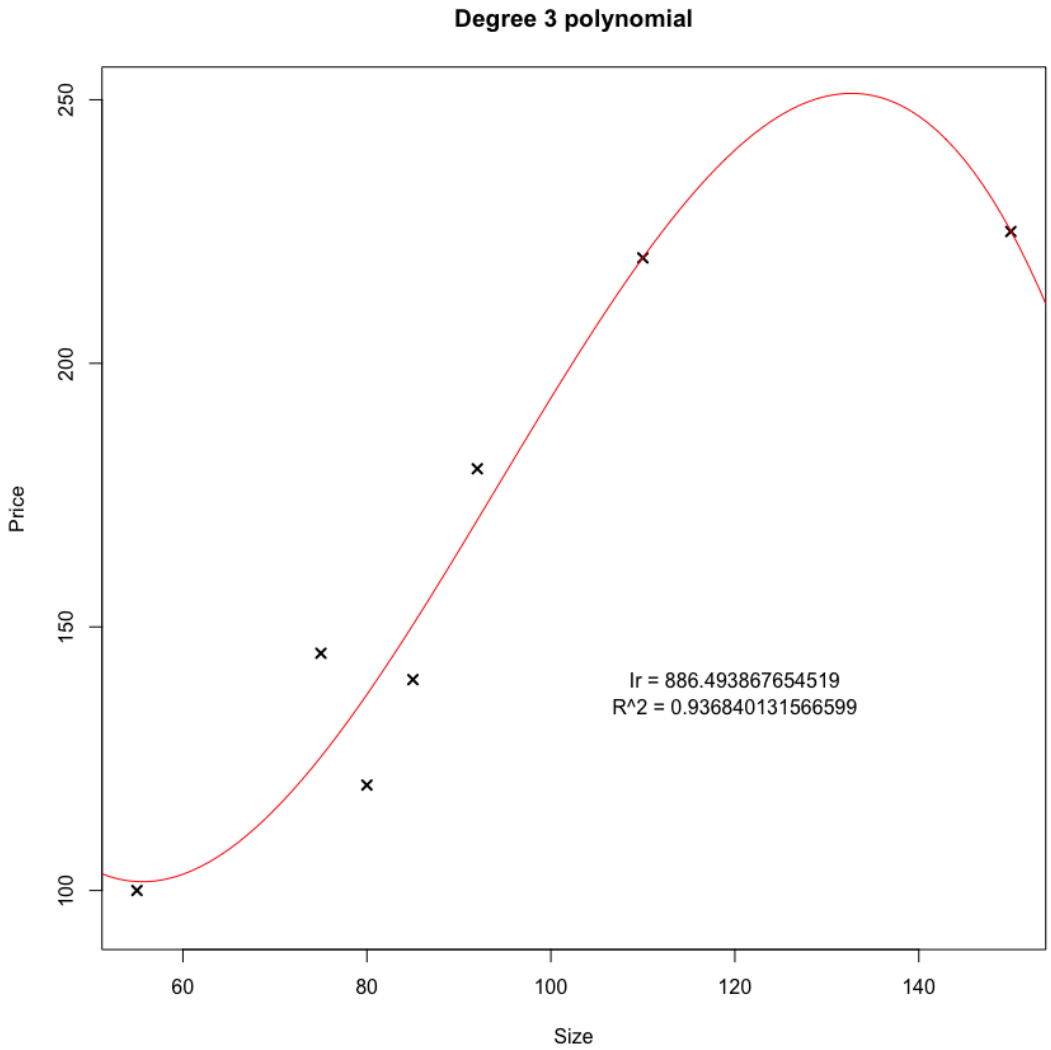
je_r et R^2 ne sont pas de bons indicateurs

Taille	$Taille^2$	Prix
80	6400	120
92	8464	180
75	5625	145
110	12100	220
85	7225	140
150	22500	225
55	3025	100



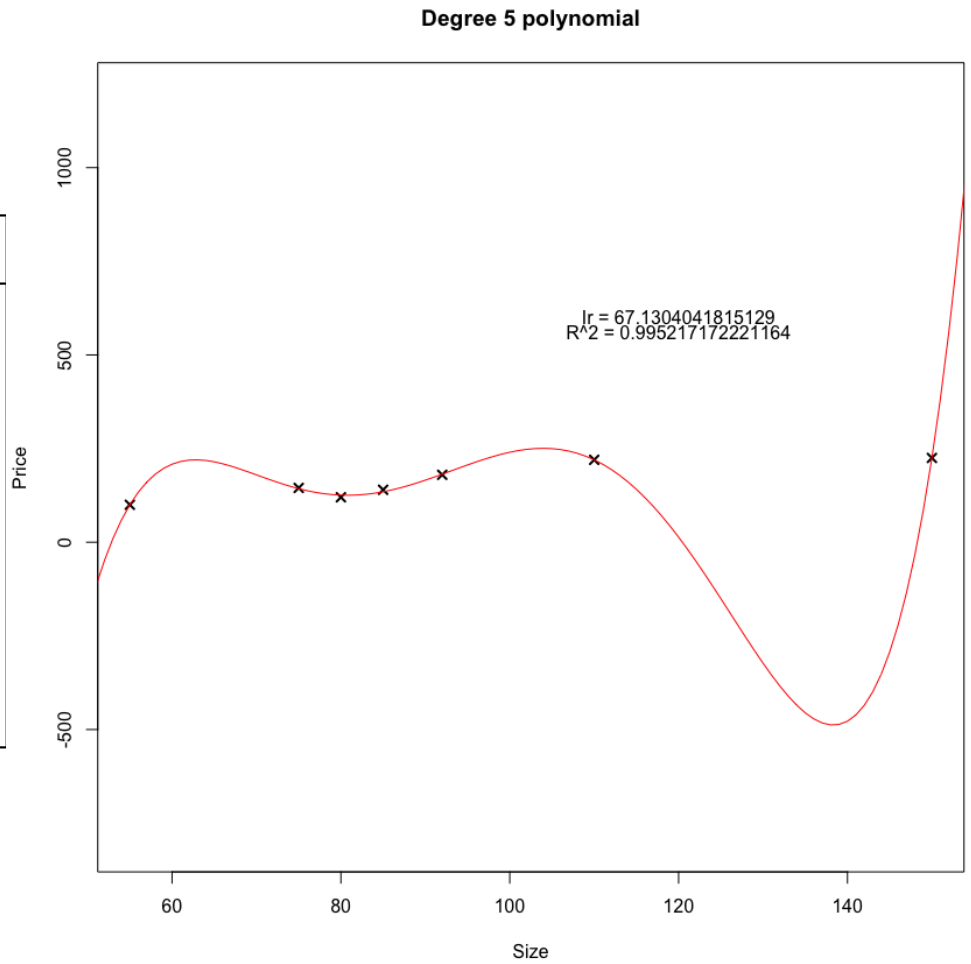
je_r et R^2 ne sont pas de bons indicateurs

Taille	$Taille^2$	$Taille^3$	Prix
80	6400		120
92	8464		180
75	5625		145
110	12100		220
85	7225		140
150	22500		225
55	3025		100



je_r et R^2 ne sont pas de bons indicateurs

Taille	$Taille^2$	$Taille^3$	$Taille^4$	$Taille^5$
80	6400			
92	8464			
75	5625			
110	12100			
85	7225			
150	22500			
55	3025			



Erreur de généralisation d'un modèle

Définition : c'est l'erreur moyenne qu'un modèle ferait en prédisant de nouveaux individus.

Problème : Il est impossible de le savoir exactement car nous n'avons pas de nouveaux individus...

→ nous allons essayer d'avoir une estimation juste de cette valeur. 2

méthodes principales :

1. Séparation Train/Test
2. Validation croisée K-fold

Séparation Train/Test

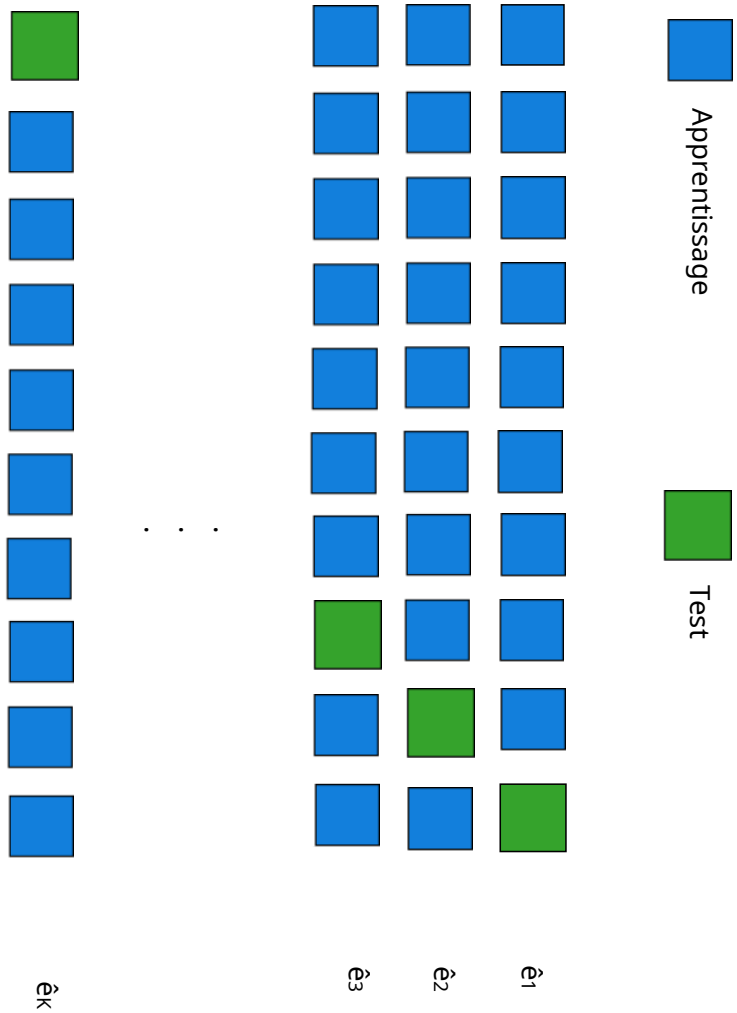
Individus	X_1	...	X_j	...	X_p	Oui
1	x_{11}	...	x_{1j}	...	x_{1p}	et_1
	...		...		...	
je	x_{je1}	...	x_{ij}	...	$x_{adresse\ IP}$	et_{je}
je +1	$x_{je+1,1}$	...	$x_{je+1,j}$	...	$x_{je+1,p}$	et_{je+1}
	...		...		...	
n	x_{n1}	...	$x_{New\ Jersey}$	...	x_{np}	et_n

Les individus dansbleusont habitués àapprendreun modèle (ou différents modèles)
Le modèle est utilisé pour prédireetpour tous les individus dansvert L'erreur de généralisation est estimée par une erreur quadratique moyenne de ces prédictions.

Inconvénient ?

Validation croisée K-fold

Principe : le jeu de données est divisé en K



L'estimation de l'erreur de généralisation est la moyenne des K erreurs de mesure.