# Introduction Stats Econométrie

## Régression multivariée

**marco.cuturi@ensae.fr**

# Régression: fondamentaux

- Même sujet peut être étudié différemmet: statistiques, algèbre linéaire, AI... *etc.*.
- La régression linéaire est directement relié au récentes idées sur la **parsimonie**
  - Lasso (1996$\rightarrow$)
  - SVM régression (1996$\rightarrow$)
  - Compressed Sensing (2002$\rightarrow$)

# Une des tâches les plus fondamentales: la régression

Données: plusieurs observations du même type

- Base de données $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$.
- Chaque observation $\mathbf{x}_j$ est un vecteur de mesures $\mathbf{x}_j = \begin{bmatrix} x_{\mathbf{1},j} \\ x_{\mathbf{2},j} \\ \vdots \\ x_{\boldsymbol{d},j} \end{bmatrix}$
- Chaque mesure $x_{i,j}$ $(1 \leq i \leq d)$ d'une observation $\mathbf{x}_j$ est un nombre.

# Une des tâches les plus fondamentales: la régression

Cette base de données peut s'interpréter comme une **matrice** $\mathbb{R}^{d\times N}$

$$\{\mathbf{x}^1, \cdots, \mathbf{x}^N\} \Longleftrightarrow \begin{bmatrix} \boldsymbol{x_{1,1}} & \boldsymbol{x_{1,2}} & \cdots & \boldsymbol{x_{1,N}} \\ \boldsymbol{x_{2,1}} & \boldsymbol{x_{2,2}} & \cdots & \boldsymbol{x_{2,N}} \\ \boldsymbol{x_{3,1}} & \boldsymbol{x_{3,2}} & \cdots & \boldsymbol{x_{3,N}} \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{x_{d,1}} & \boldsymbol{x_{d,2}} & \cdots & \boldsymbol{x_{d,N}} \end{bmatrix}$$

# Exemples

$$\text{Credit card holder } \mathbf{x}_j = \begin{bmatrix} \text{Income} \\ \text{Age} \\ \vdots \\ \text{Work history (months)} \\ \text{Family} \\ \#\ \text{Credit Incidents} \end{bmatrix}$$

$$\text{Patient } \mathbf{x}_j = \begin{bmatrix} \text{height} \\ \text{weight} \\ \vdots \\ \#\ \text{minutes exercise/week} \\ \text{LDL cholesterol} \\ \text{HDL cholesterol} \end{bmatrix}$$

$$\text{Blog } \mathbf{x}_j = \begin{bmatrix} \text{avg. pages view/month} \\ \#\ \text{posts} \\ \vdots \\ \text{avg. }\#\text{ comments/month} \\ \text{revenue from ads/month} \end{bmatrix}$$

# Au sein de ces variables...

- Certaines peuvent être **faciles à mesurer**, d'autres **plus difficiles**.
- Certaines variables peuvent avoir un effet **causal** sur les autres.

- En régression , les $d$ charactéristiques (ou variables) sont séparées entre:
  - $k$ **régresseurs** (ou **variables prédictives, indépendantes**)
  - $d-k$ **variables dépendantes** (ou **prédites**).

# Régression

- En régression , les $d$ charactéristiques (ou variables) sont séparées entre:
  - $k$ **régresseurs** (ou **variables prédictives, indépendantes**)
  - $d-k$ **variables dépendantes** (ou **prédites**).

***deviner*** les **variables dépendantes**(**chères à mesurer**) grâce aux **régresseurs**.

# Le problème de la Régression

- Etant donnés,
  - Une BDD $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\} \Longleftrightarrow X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots x_{1,N} \\ x_{2,1} & x_{2,2} & \cdots x_{2,N} \\ x_{3,1} & x_{3,2} & \cdots x_{3,N} \\ \vdots & \vdots & \vdots \\ x_{d,1} & x_{d,2} & \cdots x_{d,N} \end{bmatrix}$

# Le problème de la Régression

- Etant donnés,
  - Une BDD $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\} \Longleftrightarrow X = \begin{bmatrix} \boldsymbol{x_{1,1}} & \boldsymbol{x_{1,2}} & \cdots \boldsymbol{x_{1,N}} \\ \boldsymbol{x_{2,1}} & \boldsymbol{x_{2,2}} & \cdots \boldsymbol{x_{2,N}} \\ \boldsymbol{x_{3,1}} & \boldsymbol{x_{3,2}} & \cdots \boldsymbol{x_{3,N}} \\ \vdots & \vdots & \vdots \\ \boldsymbol{x_{d-1,1}} & \boldsymbol{x_{d-1,2}} & \cdots \boldsymbol{x_{d-1,N}} \\ \boldsymbol{x_{d,1}} & \boldsymbol{x_{d,2}} & \cdots \boldsymbol{x_{d,N}} \end{bmatrix}$

  - Un ensemble de $k$ **régresseurs Reg** $\subset \{1, \cdots, d\}$
  - Un ensemble de $d-k$ **variables dépendantes Res** $\subset \{1, \cdots, d\}$

# Le problème de la Régression

- Régression = **construire une fonction** $f : \mathbb{R}^{\mathbf{k}} \to \mathbb{R}^{\mathbf{d\text{-}k}}$ telle que,

$$\forall \mathbf{x}, f((\boldsymbol{x_i})_{\boldsymbol{i} \in \text{Reg}}) \approx (\boldsymbol{x_k})_{\boldsymbol{k} \in \text{Res}}.$$

- *e.g.* if $d = 6$, $k = 4$, **Reg**$= \{1, 2, 3, 4\}$, **Res**$= \{5, 6\}$ nous rechercherons une fonction $f : \mathbb{R}^4 \to \mathbb{R}^2$,

$$f(\boldsymbol{x_1, x_2, x_3, x_4}) \approx (\boldsymbol{x_5, x_6})$$

# Plus d'Exemples

$$\text{Credit card holder } \mathbf{x}_j = \begin{bmatrix} \text{Income} \\ \text{Age} \\ \vdots \\ \text{Work history (months)} \\ \text{Family} \\ \#\ \text{Credit Incidents} \end{bmatrix}$$

$$\text{Patient } \mathbf{x}_j = \begin{bmatrix} \text{height} \\ \text{weight} \\ \vdots \\ \#\ \text{minutes exercise/week} \\ \text{LDL cholesterol} \\ \text{HDL cholesterol} \end{bmatrix}$$

$$\text{Blog } \mathbf{x}_j = \begin{bmatrix} \text{avg. pages view/month} \\ \#\ \text{posts} \\ \vdots \\ \text{avg. } \#\ \text{comments/month} \\ \text{revenue from ads/month} \end{bmatrix}$$

# Plus d'Exemples

$$\text{Credit card holder } \mathbf{x}_j = \begin{bmatrix} \text{Income} \\ \text{Age} \\ \vdots \\ \text{Work history (months)} \\ \text{Family} \\ \#\ \text{Credit Incidents} \end{bmatrix}$$

$$\text{Patient } \mathbf{x}_j = \begin{bmatrix} \text{height} \\ \text{weight} \\ \vdots \\ \#\ \text{minutes exercise/week} \\ \text{LDL cholesterol} \\ \text{HDL cholesterol} \end{bmatrix}$$

$$\text{Blog } \mathbf{x}_j = \begin{bmatrix} \text{avg. pages view/month} \\ \#\ \text{posts} \\ \vdots \\ \text{avg. } \#\ \text{comments/month} \\ \text{revenue from ads/month} \end{bmatrix}$$

# Dans les slides suivants...

Nous ne considérons que des tâches avec **une** variable **dépendante**

- Toutes les autres seront des **régresseurs**.
- Nous renommons la variable **dépendante y** et réassignons les noms $x_1, \cdots, x_d$ pour les **régresseurs**
- prédire plus qu'une variable? extension assez directe.

# Dans les slides suivants...

$\boxed{\text{Nous assumerons que } \mathbf{y} \text{ prend des valeurs } \mathbf{continues}.}$

- Quand **y** prend des valeurs discrètes, notamment binaires $\{0, 1\}$, d'autres approches sont envisagées.
- Pourtant... **binary** $\subset$ **real** : la régression "marche" sur des valeurs discrètes,
- Mais **real** $\not\subseteq$ **binary**... on parle alors de classification (HS).

# Exemple: Apartements


| 画像 | 路線名/駅名 住所 | バス 徒歩 | 賃料 管理費等 | 敷金または保証金 礼金(敷引) | 間取り 専有面積 | 築年月 (築年数) | 選択 |
|---|---|---|---|---|---|---|---|
| NEW! | 京阪鴨東線/出町柳 京都市左京区田中大堰町 間取り図 写真 | ー 5分 | 4.20万円 3,000円 | 5万円 5万円(ー) | 1R 16.00m² | '89/09 (築22年) | |
| | 叡山本線/修学院 京都市左京区山端滝ケ鼻町 間取り図 写真 | ー 7分 | 4.30万円 2,000円 | 5万円 なし(なし) | 1K 20.44m² | '95/03 (築17年) | |
| | 叡山本線/修学院 京都市左京区山端滝ケ鼻町 間取り図 写真 | ー 7分 | 4.30万円 2,000円 | 5万円 なし(なし) | 1K 20.44m² | '95/03 (築17年) | |
| | 叡山本線/一乗寺 京都市左京区一乗寺梅ノ木町 間取り図 写真 | ー 6分 | 4.30万円 2,000円 | 5万円 なし(ー) | 1K 20.00m² | '88/03 (築24年) | |
| | 京阪鴨東線/出町柳 京都市左京区吉田上阿達町 間取り図 写真 | ー 7分 | 4.30万円 2,000円 | 5万円 5万円(なし) | 1K 19.02m² | '85/02 (築27年) | |
| | 烏丸線/今出川 京都市上京区柳図子町 間取り図 写真 | ー 3分 | 4.30万円 2,000円 | なし 10万円(ー) | 1R 17.70m² | '84/10 (築27年) | |
| NEW! | 烏丸線/今出川 京都市上京区北小路室町 間取り図 写真 | ー 2分 | 4.50万円 なし | 5万円 5万円(ー) | 1K 16.00m² | '94/03 (築18年) | |
| | 京阪鴨東線/出町柳 京都市左京区田中下柳町 間取り図 写真 | ー 3分 | 4.30万円 2,500円 | 5万円 8万円(ー) | 1K 20.00m² | '85/03 (築27年) | |
| | 叡山本線/修学院 京都市左京区高野泉町 間取り図 写真 | ー 5分 | 4.05万円 5,500円 | 6万円 5万円(なし) | 1K 14.58m² | '79/01 (築33年) | |

Information sur 285 apartements au Japon, à Kyoto.

Source: http://realestate.yahoo.co.jp/

# Exemple: Apartements


現在の条件に合う物件数 1,226件中161〜180件を表示しています。 前へ 5 6 7 8 9 10 11 12 13 14 次へ

絞り込み条件をリセット

選択路線
叡山本線
出町柳(824)
元田中(380) 茶山(388)
一乗寺(279)
修学院(187)
路線を選びなおす 駅を選びなおす

基本条件
賃料 選び方を変える
安 8万円 高
下限なし〜上限なし
管理費・共益費込み(1,226)
礼金なし(266)
敷金・保証金なし(224)

間取り
1R(102) 1K/1DK(778)
1LDK(66) 2K/2DK(73)
2LDK(99) 3K/3DK(22)
3LDK(78) 4K/4DK(0)
4LDK以上(8)

最寄り駅からの時間(徒歩)
1分以内(63) 5分以内(466)
7分以内(706) 10分以内(902)
15分以内(1,076)
指定なし(1,226)

一覧表示 間取り表示 すべてにチェック チェックした物件をまとめて 詳細表示 お問い合わせ

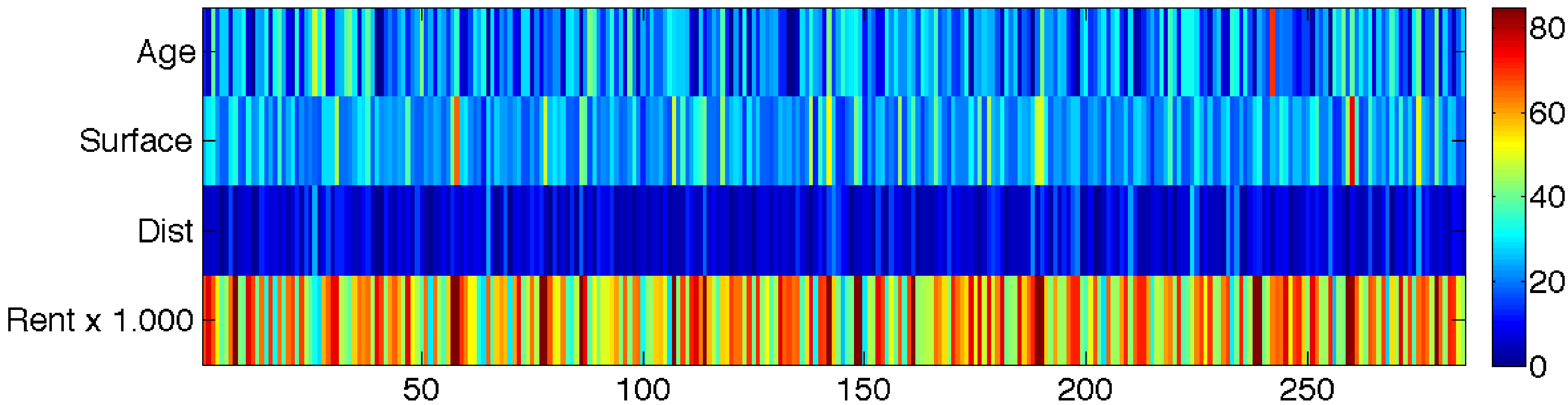
| 画像 | 路線名/駅名 住所 | バス 徒歩 | 賃料 管理費等 | 敷金または保証金 礼金(敷引) | 間取り 専有面積 | 築年月 (築年数) | 選択 |
|---|---|---|---|---|---|---|---|
| NEW! | 京阪鴨東線/出町柳 京都市左京区田中大堰町 間取り図 写真 | ― 5分 | 4.20万円 3,000円 | 5万円 5万円(―) | 1R 16.00m² | '89/09 (築22年) | |
| | 叡山本線/修学院 京都市左京区山端滝ケ鼻町 間取り図 写真 | ― 7分 | 4.30万円 2,000円 | 5万円 なし(なし) | 1K 20.44m² | '95/03 (築17年) | |
| | 叡山本線/修学院 京都市左京区山端滝ケ鼻町 間取り図 写真 | ― 7分 | 4.30万円 2,000円 | 5万円 なし(なし) | 1K 20.44m² | '95/03 (築17年) | |
| | 叡山本線/一乗寺 京都市左京区一乗寺梅ノ木町 間取り図 写真 | ― 6分 | 4.30万円 2,000円 | 5万円 なし(―) | 1K 20.00m² | '88/03 (築24年) | |
| | 京阪鴨東線/出町柳 京都市左京区吉田上阿達町 間取り図 写真 | ― 7分 | 4.30万円 2,000円 | 5万円 5万円(なし) | 1K 19.02m² | '85/02 (築27年) | |
| | 烏丸線/今出川 京都市上京区柳図子町 間取り図 写真 | ― 3分 | 4.30万円 2,000円 | なし 10万円(―) | 1R 17.70m² | '84/10 (築27年) | |
| NEW! | 烏丸線/今出川 京都市上京区北小路室町 間取り図 写真 | ― 2分 | 4.50万円 なし | 5万円 5万円(―) | 1K 16.00m² | '94/03 (築18年) | |
| | 京阪鴨東線/出町柳 京都市左京区田中下柳町 間取り図 写真 | ― 3分 | 4.30万円 2,500円 | 5万円 8万円(―) | 1K 20.00m² | '85/03 (築27年) | |
| | 叡山本線/修学院 京都市左京区高野泉町 間取り図 写真 | ― 5分 | 4.05万円 5,500円 | 6万円 5万円(なし) | 1K 14.58m² | '79/01 (築33年) | |

Information sur 285 apartements au Japon, à Kyoto.

Retenu 4 variables: **Surface**, **Loyer**, **Age du batiment**, **distance / station**.

# What does the matrix look like?

```
imagecs(H); colorbar;
```

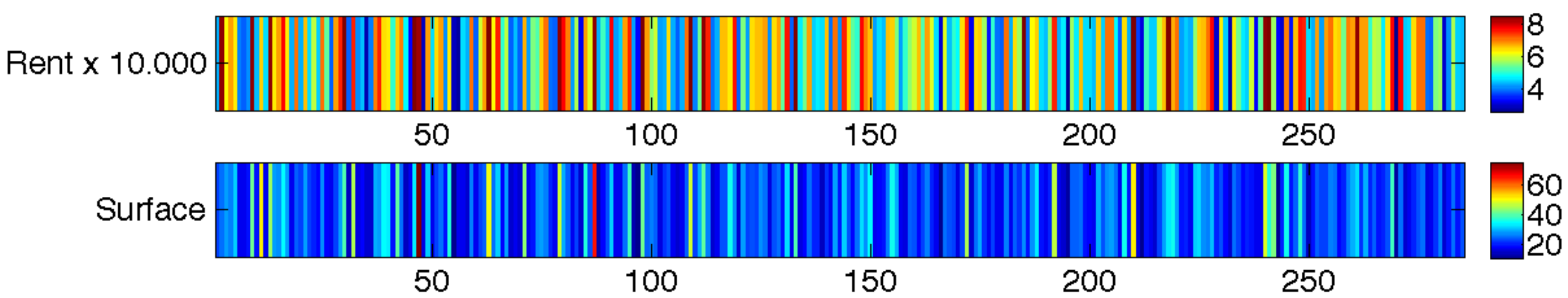



285 columns, 4 lines.

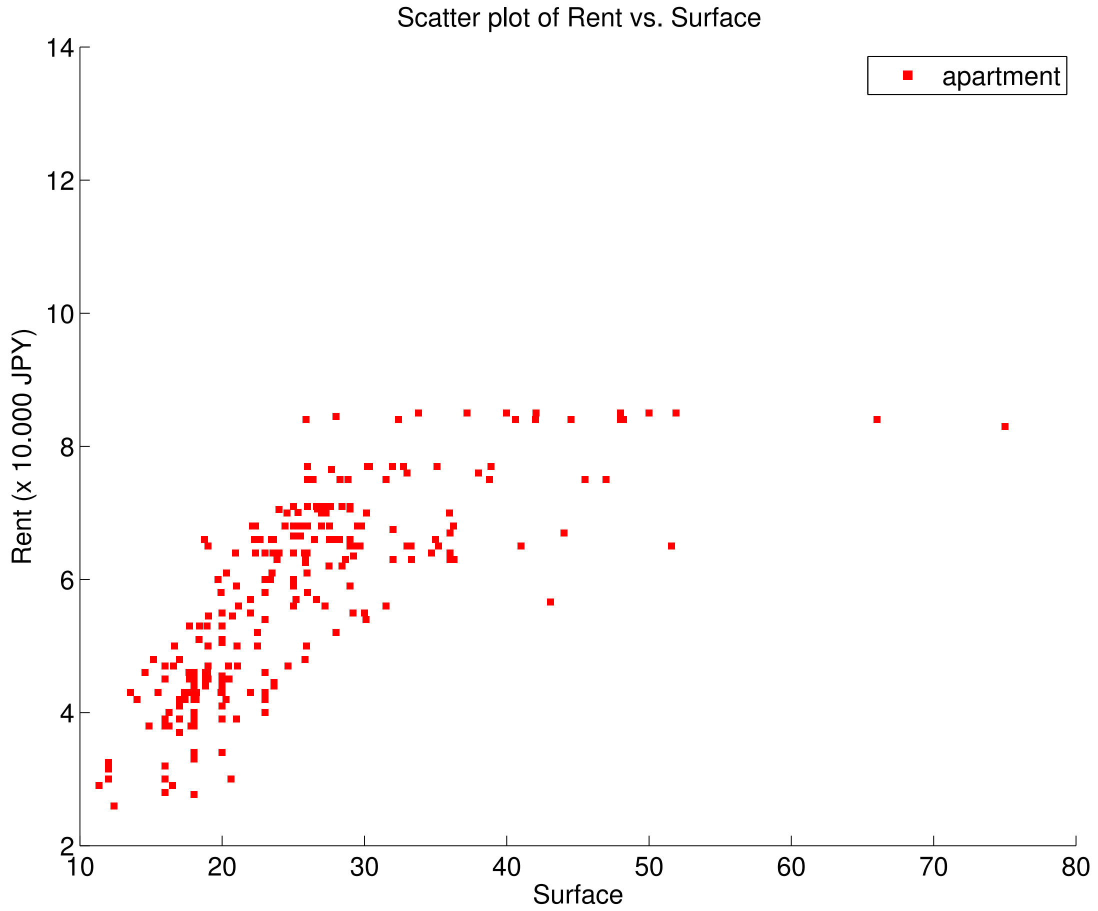
Chaque colonne représente un apartement.

| Nous ferons la **régression** du loyer à partir de l'âge, surface and distance |
|---|

# Régression: une variable vs. une autre

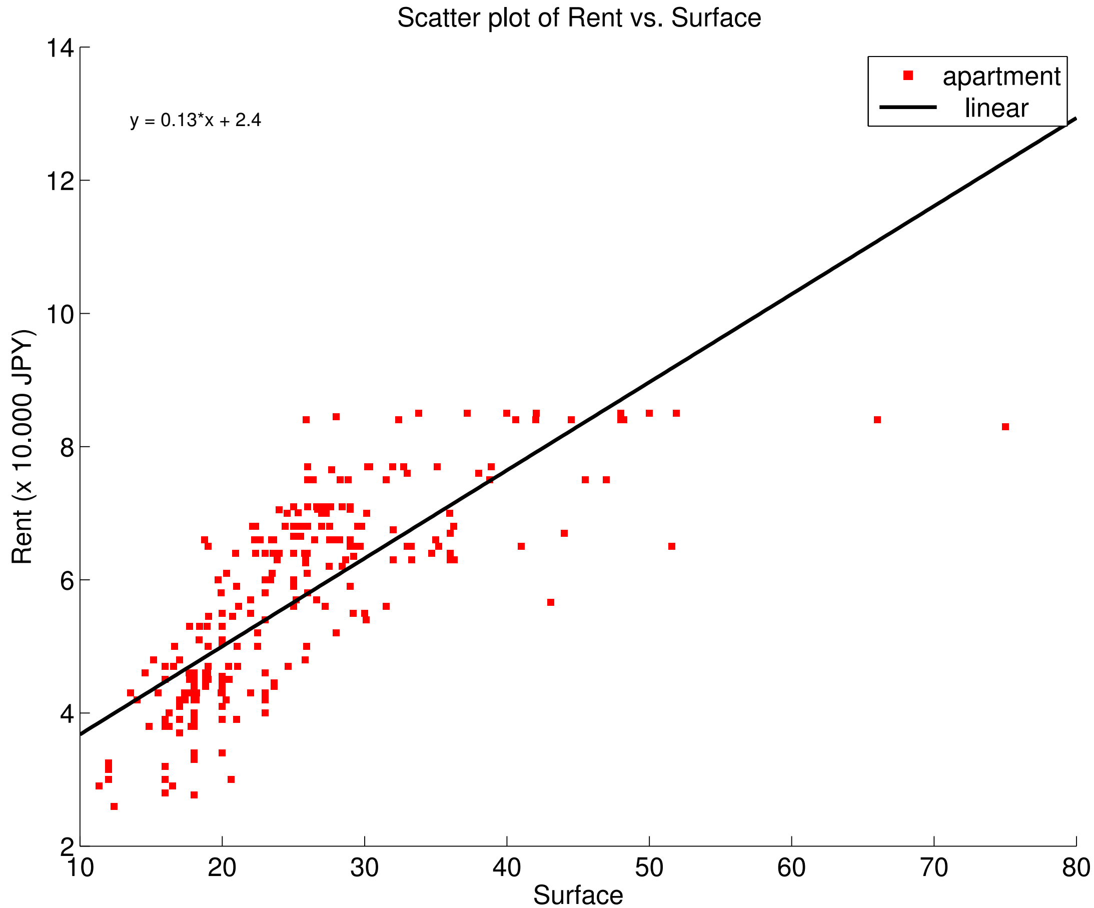


# Loyer vs. Surface



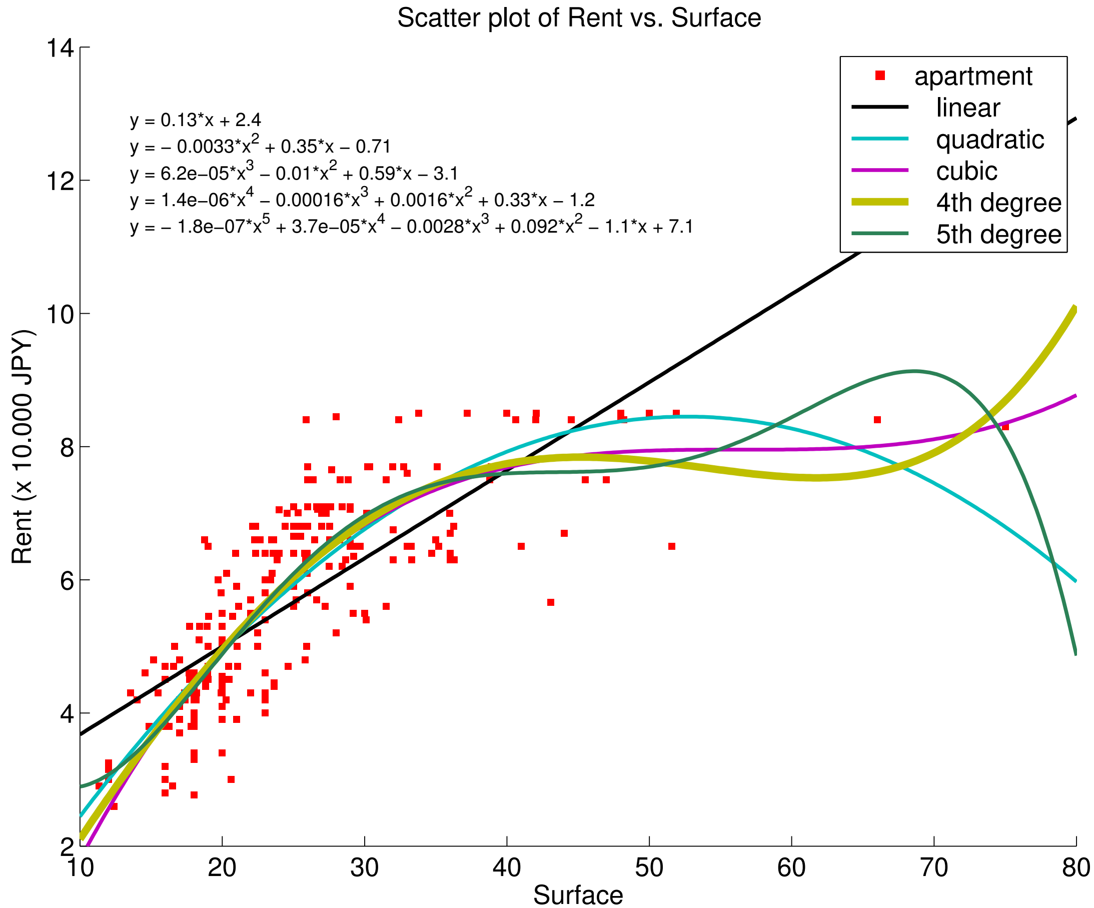

Remarquez que le jeu de donnes est censuré au dessus de 85.000 JPY

# Loyer vs. Surface




Si nous utilisons un outil logiciel, nous obtenons l'approximation **y** $= 0.13x + 2.4$

# Loyer vs. Surface




Nous pouvons utiliser des polynômes de degré supérieur....

# Outil logiciel de "curve fitting"

- Tous ces logiciels utilisent le critère des **moindres carrés** e.g

$$\min_{\boldsymbol{f} \in \mathcal{F}} \sum_{j=1}^{N} (\boldsymbol{y_j} - \boldsymbol{f}(\boldsymbol{x_j}))^2$$

  où $\mathcal{F}$ est une **classe de functions**

- Les logiciels courants considèrent plusieurs classes for $\mathcal{F}$.

# Behind the curve fitting tool

- Tous ces logiciels utilisent le critère des **moindres carrés** e.g

$$\min_{\boldsymbol{f}\in\mathcal{F}} \sum_{j=1}^{N} (\boldsymbol{y_j} - \boldsymbol{f}(\boldsymbol{x_j}))^2$$

  où $\mathcal{F}$ est une **classe de functions**

- Les logiciels courants considèrent plusieurs classes for $\mathcal{F}$.

  ◦ **Linéaire** $\min_{b,a_1\in\mathbb{R}} \sum_{j=1}^{N} \left(\boldsymbol{y_j} - (\boldsymbol{b} + \boldsymbol{a_1}\boldsymbol{x_j})\right)^2$

# Behind the curve fitting tool

- Tous ces logiciels utilisent le critère des **moindres carrés** e.g

$$\min_{f \in \mathcal{F}} \sum_{j=1}^{N} (y_j - f(x_j))^2$$

où $\mathcal{F}$ est une **classe de functions**

- Les logiciels courants considèrent plusieurs classes for $\mathcal{F}$.
  - **Linéaire** $\min_{b, a_1 \in \mathbb{R}} \sum_{j=1}^{N} (y_j - (b + a_1 x_j))^2$
  - **Quadratique** $\min_{b, a_1, a_2 \in \mathbb{R}} \sum_{j=1}^{N} \left(y_j - (b + a_1 x_j + a_2 x_j^2)\right)^2$

# Behind the curve fitting tool

- Tous ces logiciels utilisent le critère des **moindres carrés** e.g

$$\min_{f \in \mathcal{F}} \sum_{j=1}^{N} (y_j - f(x_j))^2$$

où $\mathcal{F}$ est une **classe de functions**

- Les logiciels courants considèrent plusieurs classes for $\mathcal{F}$.
  - **Linéaire** $\min_{b, a_1 \in \mathbb{R}} \sum_{j=1}^{N} (y_j - (b + a_1 x_j))^2$
  - **Quadratique** $\min_{b, a_1, a_2 \in \mathbb{R}} \sum_{j=1}^{N} \left(y_j - (b + a_1 x_j + a_2 x_j^2)\right)^2$
  - **Cubique** $\min_{b, a_1, a_2, a_3 \in \mathbb{R}} \sum_{j=1}^{N} \left(y_j - (b + a_1 x_j + a_2 x_j^2 + a_3 x_j^3)\right)^2$
  - *etc.*

# Comment résoudre ce problème? le cas linéaire

- Jetons un oeil à la fonction

$$(a,b) \mapsto \sum_{j=1}^{N} \left(\boldsymbol{y_j} - (\boldsymbol{b} + \boldsymbol{a}x_j)\right)^2.$$

- Using the notations

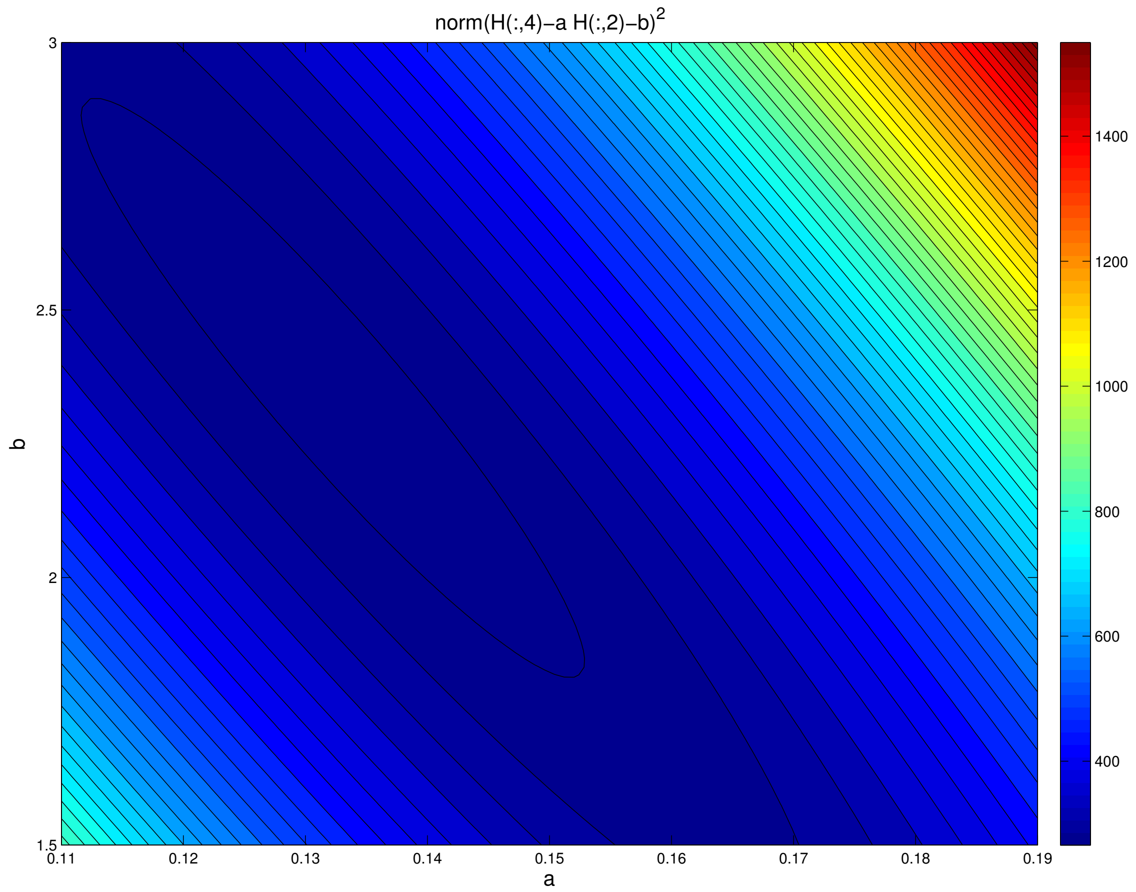
$$\begin{array}{rl} \text{Loyer} & Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix} \\ \text{Surface} & X = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \\ \text{Constante} & \mathbf{1}_N = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \end{array}$$

Nous avons

$$\sum_{j=1}^{N} \left(\boldsymbol{y_j} - (\boldsymbol{a}\boldsymbol{x_j} + \boldsymbol{b})\right)^2 = \|\boldsymbol{Y} - \boldsymbol{a}\boldsymbol{X} - \boldsymbol{b}\mathbf{1}_N\|^2$$

# Contour plot de $(a, b) \rightarrow \|Y - aX - b\mathbf{1}_N\|^2$

- Comme nous n'avons que 2 paramètres, nous pouvons produire un contour plot




- Ceci valide la réponse $\mathbf{y} = 0.13x + 2.4$. Comment l'obtenir?

# Un peu d'algèbre linéaire

- We define the function $L$ as

$$L : (a, b) \mapsto \sum_{j=1}^{N} \left(\boldsymbol{y_j} - (\boldsymbol{b} + \boldsymbol{a}x_j)\right)^2$$

- Les dérivées partielles de $L$ peuvent se calculer.

$$\frac{\partial L}{\partial a} = -2 \sum_{j=1}^{N} \left(\boldsymbol{y_j} - (\boldsymbol{b} + \boldsymbol{a}x_j)\right) \boldsymbol{x_j}$$

$$\frac{\partial L}{\partial b} = -2 \sum_{j=1}^{N} \boldsymbol{y_j} - (\boldsymbol{b} + \boldsymbol{a}x_j)$$

- Tout minimum $(a^\star, b^\star)$ de $L$ est un point selle.

# Un peu d'algèbre linéaire

- The dérivées partielles de $L$ doivent s'annuler en $(a^\star, b^\star)$

$$\frac{\partial L}{\partial a} = 2\left(\boldsymbol{a}\sum \boldsymbol{x_j^2} + \boldsymbol{b}\sum \boldsymbol{x_j} - \sum y_j x_j\right)$$
$$\frac{\partial L}{\partial b} = 2\left(N\boldsymbol{b} - \sum y_j + \boldsymbol{a}\sum x_j\right)$$

- Ainsi, $(a^\star, b^\star)$ **doit satisfaire** le système linéaire

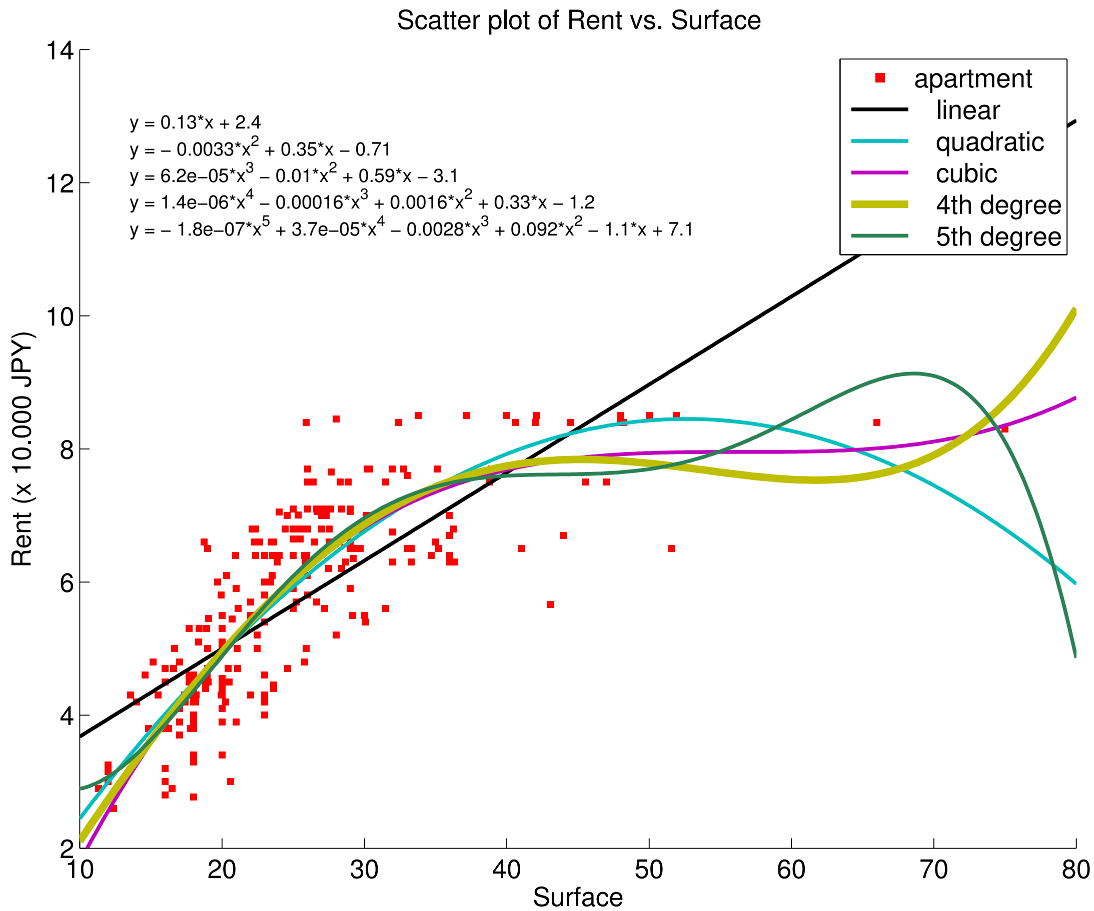
$$0 = a^\star \sum \boldsymbol{x_j^2} + b^\star \sum \boldsymbol{x_j} - \sum y_j \boldsymbol{x_j}$$
$$0 = Nb^\star - \sum y_j + a^\star \sum \boldsymbol{x_j}$$

- Soit:

$$\begin{bmatrix} a^\star \\ b^\star \end{bmatrix} = \begin{bmatrix} \sum \boldsymbol{x_j^2} & \sum \boldsymbol{x_j} \\ \sum \boldsymbol{x_j} & N \end{bmatrix}^{-1} \begin{bmatrix} \sum y_j x_j \\ \sum y_j \end{bmatrix}$$

- `ans = 0.132248772789152 2.354203561671262`

# Loyer vs. Surface




Nous avons compris le cas linéaire. Quid du cas quadratique?

# Cas quadratique

$$\textbf{Quadratique} \quad \min_{b,a_1,a_2\in\mathbb{R}} \sum_{j=1}^{N} \left(\boldsymbol{y_j} - (\boldsymbol{b} + \boldsymbol{a_1 x_j} + \boldsymbol{a_2 x_j^2})\right)^2$$

- même idée... on défini

$$L : (a_1, a_2, b) \mapsto \sum_{j=1}^{N} \left(\boldsymbol{y_j} - \left(\boldsymbol{b} + \boldsymbol{a_1 x_j} + \boldsymbol{a_2 x_j^2}\right)\right)^2$$

- regardons maintenant les dérivées de cet objectif...

$$\frac{\partial L}{\partial a_2} = -2\sum_{j=1}^{N} \left(\boldsymbol{y_j} - \left(\boldsymbol{b} + \boldsymbol{a_1 x_j} + \boldsymbol{a_2 x_j^2}\right)\right) \boldsymbol{x_j^2}$$

$$\frac{\partial L}{\partial a_1} = -2\sum_{j=1}^{N} \left(\boldsymbol{y_j} - \left(\boldsymbol{b} + \boldsymbol{a_1 x_j} + \boldsymbol{a_2 x_j^2}\right)\right) \boldsymbol{x_j}$$

$$\frac{\partial L}{\partial b} = -2\sum_{j=1}^{N} \left(\boldsymbol{y_j} - \left(\boldsymbol{b} + \boldsymbol{a_1 x_j} + \boldsymbol{a_2 x_j^2}\right)\right)$$

## Cas quadratique

- Equations pour un point selle:

$$0 = \sum_{j=1}^{N} \left( \boldsymbol{y_j} - \left( b^\star + a_1^\star \boldsymbol{x_j} + a_2^\star \boldsymbol{x_j^2} \right) \right) \boldsymbol{x_j^2}$$

$$0 = \sum_{j=1}^{N} \left( \boldsymbol{y_j} - \left( b^\star + a_1^\star \boldsymbol{x_j} + a_2^\star \boldsymbol{x_j^2} \right) \right) \boldsymbol{x_j}$$

$$0 = \sum_{j=1}^{N} \left( \boldsymbol{y_j} - \left( b^\star + a_1^\star \boldsymbol{x_j} + a_2^\star \boldsymbol{x_j^2} \right) \right)$$

$$\begin{bmatrix} a_2^\star \\ a_1^\star \\ b^\star \end{bmatrix} = \begin{bmatrix} \sum \boldsymbol{x_j^4} & \sum \boldsymbol{x_j^3} & \sum \boldsymbol{x_j^2} \\ \sum \boldsymbol{x_j^3} & \sum \boldsymbol{x_j^2} & \sum \boldsymbol{x_j} \\ \sum \boldsymbol{x_j^2} & \sum \boldsymbol{x_j} & N \end{bmatrix}^{-1} \begin{bmatrix} \sum \boldsymbol{y_j x_j^2} \\ \sum \boldsymbol{y_j x_j} \\ \sum \boldsymbol{y_j} \end{bmatrix}$$

- `ans = -0.003306463076068 0.347969105896777 -0.705157514974559`

# Polynômes d'ordre sup ´rieur

- Intuitivement, pour les polynômes d'ordre $p$ nous auronsà
  - Construire une matrice (de Toeplitz)
  - Combiner les valeurs de $y$ and $x$ pour plusieurs exposants
  - Résoude le système linéaire
- D'une manière peut-être surprenante:

| |
|---|
| Trouver le **meilleur polynôme d'ordre** $p$ avec les **moindres carrés** |
| $\Updownarrow$ |
| Résoudre un système linéaire de $p$ **variables** |

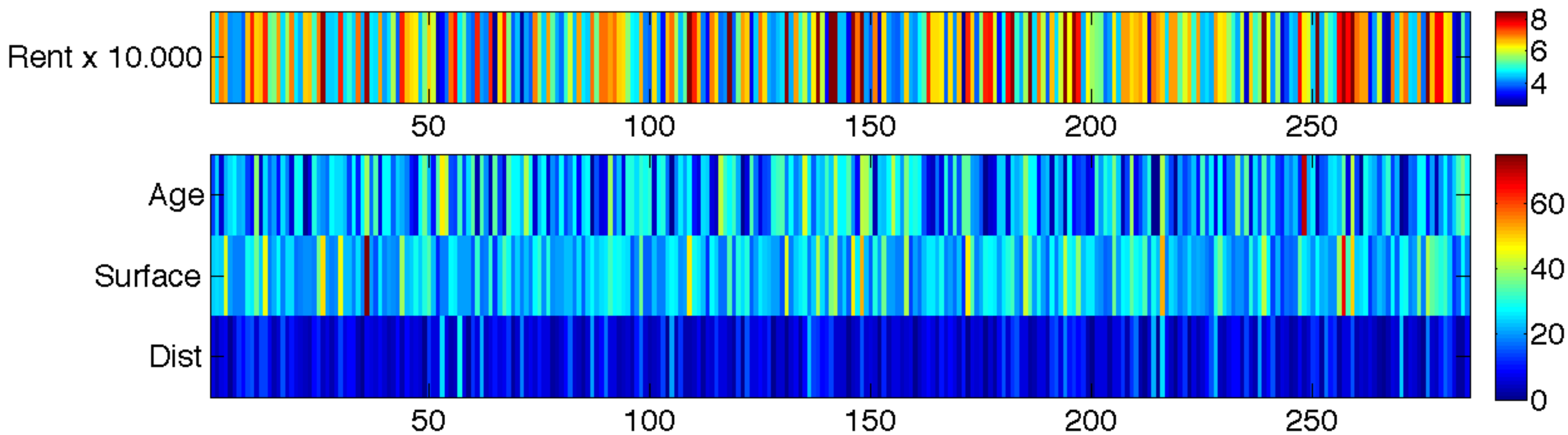
- Pas si surprenant en fait:
  - Moindre carrés: objectif de degré 2 en les coefficients.
  - Minimum $\Leftrightarrow$ point selle $\Leftrightarrow$ système de degré 1..
  - Les moindre carrés sont choisis **parce qu'** ils sont résolus via système linéaire...

# Le cas général : une vs. plusieurs variables

- Comment pouvons nous introduire plusieurs variables?

Loyer $Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}$

All other variables $X = \begin{bmatrix} \vdots & \vdots & \cdots & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix}$



# Le cas général

- Nous avons donc $d$ **régresseurs** , $1$ **variable dépendante**.
- Considérons à nouveau l'approche **linéaire**. Nous voulons une fonction $f$ de la forme

$$f(\mathbf{x}) = \boldsymbol{\alpha_0} + \boldsymbol{\alpha_1} x_1 + \boldsymbol{\alpha_2} x_2 + \cdots + \boldsymbol{\alpha_d} x_d.$$

- Nous devons déterminer $d+1$ valeurs
  - une constante $\boldsymbol{\alpha_0}$
  - des poids $\boldsymbol{1 \leq i \leq d, \alpha_i}$ pour chaque variable.

- **Moindres Carrés**:

$$L(\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_d) = \sum_{j=1}^{N} \Big( \boldsymbol{y_j} - (\boldsymbol{\alpha_0} + \boldsymbol{\alpha_1} x_{1,j} + \boldsymbol{\alpha_2} x_{2,j} + \cdots + \boldsymbol{\alpha_d} x_{d,j}) \Big)^2$$

# Le cas général

- Remarquons que

$$L(\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_d) \to \sum_{i=1}^{N} \left( y_i - \left( \alpha_0 + \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_d \end{bmatrix}^T \mathbf{x}_i \right) \right)^2 = \left\| \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_d \end{bmatrix}^T X - Y \right\|^2,$$

where

$$X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \\ \vdots & \vdots & \cdots & \vdots \end{bmatrix} \in \mathbb{R}^{d+1 \times N}$$

and

$$Y = \begin{bmatrix} y_1 & \cdots & y_N \end{bmatrix} \in \mathbb{R}^N.$$

- Nous écrivons $\alpha$ pour le vecteur de taille $d+1$ $\begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_d \end{bmatrix}$.

# Moindres Carrés - Linéaire

- En développant cette expression

$$L(\alpha) = \left(\alpha^T XX^T\alpha - 2YX^T\alpha + \|Y\|^2\right)$$

- Considérons le **gradient** de cette fonction:

$$\nabla L = 2XX^T\alpha - 2XY^T$$

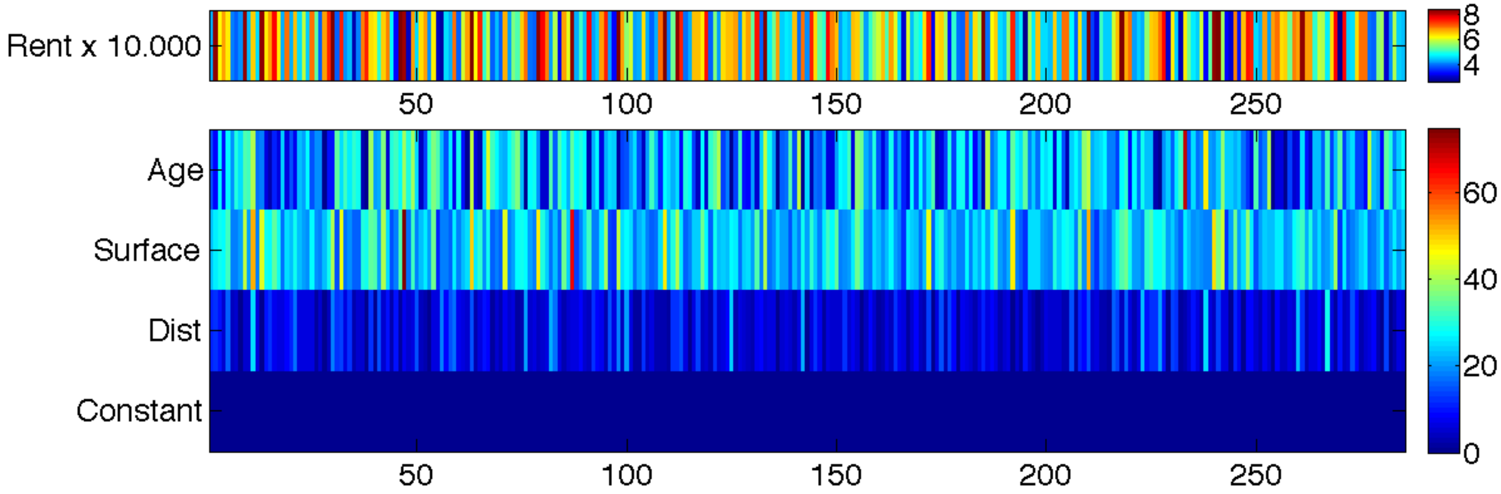
- Le gradient s'annule donc si

$$\alpha^\star = (XX^T)^{-1}XY^T$$

- $XX^T \in \mathbf{S}_+^n$: $XX^T$ est une matrice semidéfinie positive.
- Cette expression marche si $XX^T \in \mathbb{R}^{d+1}$ est **inversible**, à savoir $XX^T \in \mathbf{S}_{++}^n$.

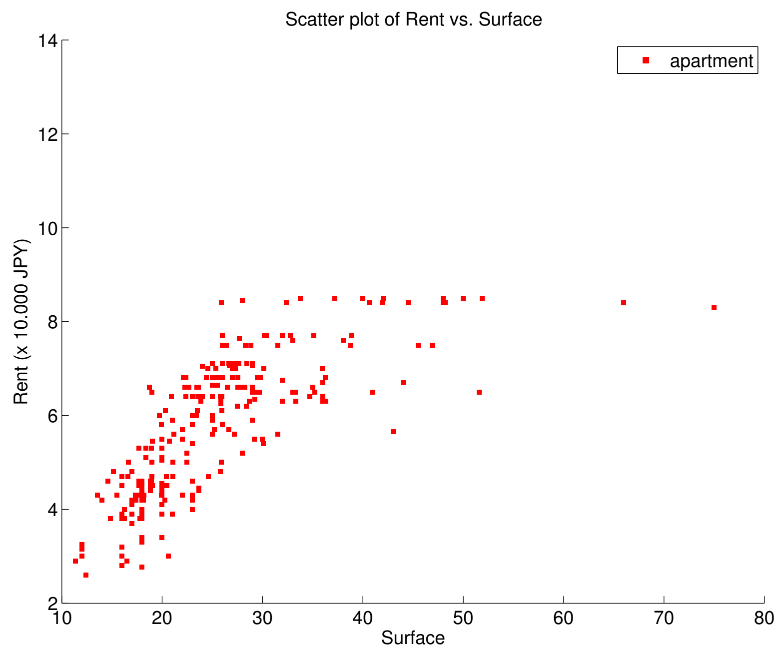
# Loyer vs. le reste

- Prenons nos données. Ajoutons une ligne de $\mathbf{1}s$



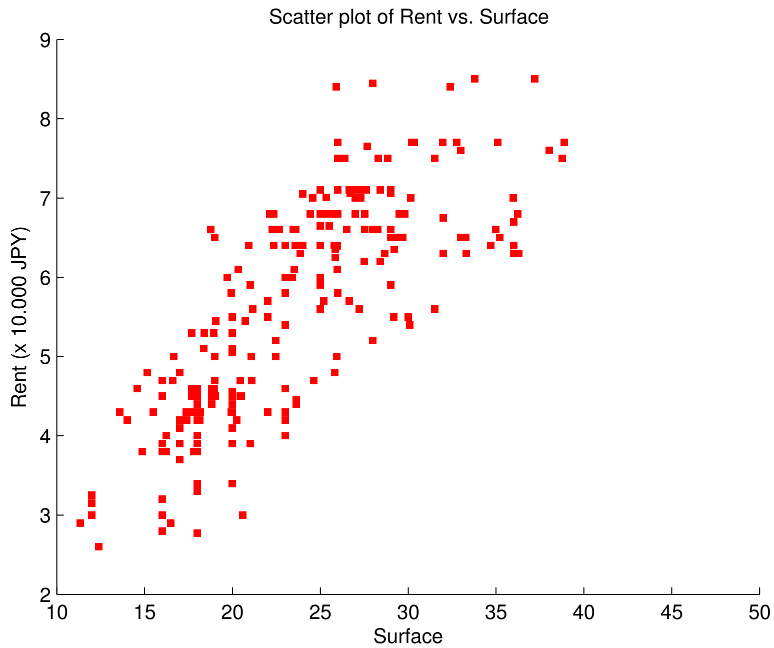

```
>> (X*X')\(X*Y')
ans =
    0.000141678721821
    0.004226687659299
   -0.012599982792209
    5.611128285287092
```

# Quel est le problème?




$$\boxed{\text{loyer} = 1.4 \text{ âge} + 42.2 \text{ surf} - 125 \text{ dist} + 56.110 \text{ JPY}}$$

# En enlevant les points aberrants? (surf$> 40$)




```
>> (X*X')\(X*Y')
ans =
   -0.049332605603095     x age
    0.163122792160298     x surface
   -0.004411580036614     x distance
    2.731204399433800     + 27.300 JPY
```

Morale: facile de tirer des conclusions fausses

# Que peut il arriver de problématique?

- Que se passe-t-il quand $d \gg n$? $(XX^T)$ n'est **plus inversible**...
  - données en génomique,
  - données en analyse d'images (beaucoup de charactéristiques)

- Que se passe-t-il quand $(XX^T)$ est **mal conditionnée?** $(\frac{\lambda_{\min}(XX^T)}{\lambda_{\max}(XX^T)} \approx 0)$?
  - Si $\lambda_{\min}(XX^T) = 1e - 10$, $\lambda_{\max}\left((XX^T)^{-1}\right) = 1e10$!!
  - Très mauvaise stabilité numérique de la solution.

- Quand $d \gg n$, nous sommes tentés de faire de la **sélection de variable**,
  - $i.e.$choisir un sous-ensemble $d'$ des $d$ variables qui suffit pour prédire **y**.
  - $i.e.$favoriser des vecteurs $\alpha$ tels que $\|\alpha\|_0 = \mathrm{card}\{\alpha_i \neq 0\}$ est petit.

# En résumé

**Régression**: trouver une relation fonctionnelle
entre une **variable dépendante** and des **prédicteurs**

# En résumé

**Régression**: trouver une relation fonctionnelle
entre une **variable dépendante** and des **prédicteurs**

**Trouver une function $\boldsymbol{f}$ telle que**

$$\forall(\mathbf{x}, \boldsymbol{y}) \text{ observables }, \boldsymbol{f}(\mathbf{x}) \approx \boldsymbol{y}$$

# En résumé

**Régression**: trouver une relation fonctionnelle
entre une **variable dépendante** and des **prédicteurs**

**Trouver une function $f$ telle que**

$$\forall(\mathbf{x}, \boldsymbol{y}) \text{ observables }, \boldsymbol{f}(\mathbf{x}) \approx \boldsymbol{y}$$

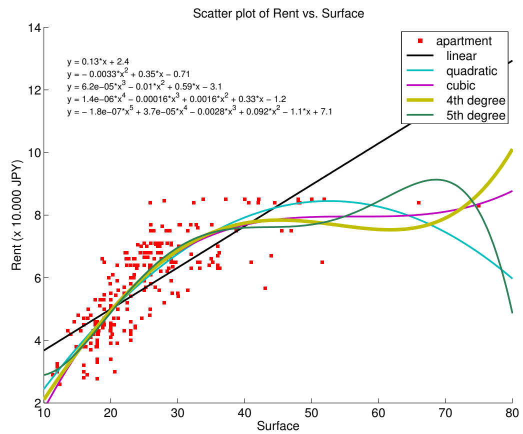
**Utiliser une BDD & le critère des moindres carrés**:

$$\min_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{N} \sum_{j=1}^{N} (\boldsymbol{y_j} - \boldsymbol{f}(\boldsymbol{x_j}))^2$$

# En résumé

**Régression**: trouver une relation fonctionnelle
entre une **variable dépendante** and des **prédicteurs**

- **pour la régression d'une variable réelle *vs.* une autre** :




  - Moindres-Carrés $L(b, a_1, \cdots, a_p)$ pour estimer **lignes**, polynômes.
  - Aboutit à un système linéaire.

$$\frac{\partial \mathbf{2}^{\text{nd}}\text{order}(b, a_1, \cdots, a_p)}{\partial a_p} = \textbf{linear} \text{ in } (b, a_1, \cdots, a_p)$$

  - En posant $\partial L / \partial a_p = 0$ nous obtenons $p + 1$ equations **linéaires** pour $p + 1$ variables.

# En résumé

**Régression**: trouver une relation fonctionnelle
entre une **variable dépendante** and des **prédicteurs**

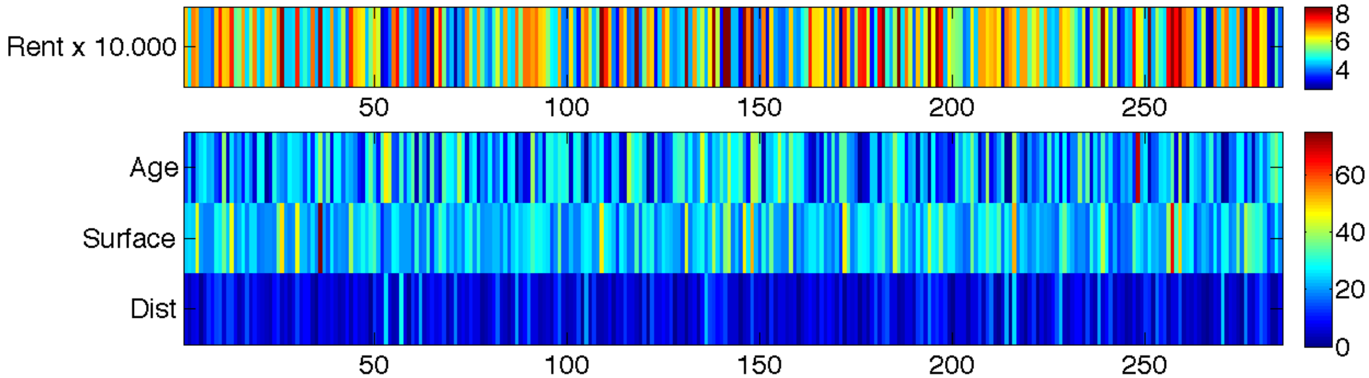
- **pour la régression d'une variable réelle vs plusieurs autres** :
  - Trouver vecteur de poids $\alpha \in \mathbb{R}^d$ tel que $(\alpha^T\mathbf{x} + \alpha_0) \approx y$.
  - On ajoute à la matrice $d \times N$ des données, une ligne de $1$ pour définir $\boldsymbol{X}$.
  - La ligne $\boldsymbol{Y}$ de valeurs **prédire**.
  - Critère des Moindres-Carrés:

$$L(\alpha) = \|\boldsymbol{Y} - \alpha^T\boldsymbol{X}\|^2 = \left(\alpha^T\boldsymbol{X}\boldsymbol{X}^T\alpha - 2\boldsymbol{Y}\boldsymbol{X}^T\alpha + \|\boldsymbol{Y}\|^2\right).$$

$$\nabla_\alpha L = 0 \quad \Rightarrow \quad \alpha^\star = \boldsymbol{(XX^T)^{-1}XY^T}$$

- Ceci marche si $\boldsymbol{X}\boldsymbol{X}^T \in \mathbb{R}^{d+1}$ est **inversible**.

# Exemple




```
>> (X*X')\(X*Y')
ans =
   -0.049332605603095     x age
    0.163122792160298     x surface
   -0.004411580036614     x distance
    2.731204399433800     + 27.300  JPY
```

# Autres interprétations

- Une perspective **statistique** sur la régression MC.
- Quelques mots sur l'utilisation de **polynômes** en plus haute dimension.
- Une perspective **géométrique**
- **co-linéarité des variables et surapprentissage**
- Quelques solutions: **techniques de régression avancées**
  - Sélection de variables
  - Régression "Ridge"
  - Lasso