Quelques éléments de statistique

B. Michel

Ecole Centrale de Nantes

Outline

Un exemple: température corporelle

Le jeu de données **Heart Rate** renseigne pour 130 individus la température corporelle, le sexe and son rythme cardiaque.

```
##     BodyTemp Gender HeartRate
## 1       96.3   male        70
## 2       96.7   male        71
## 100     98.4 female        84
## 101     98.5 female        83
## 102     98.6 female        82
## 103     98.6 female        85
```

Un exemple: température corporelle

Le jeu de données **Heart Rate** renseigne pour 130 individus la température corporelle, le sexe and son rythme cardiaque.

```
##      BodyTemp Gender HeartRate
## 1        96.3   male        70
## 2        96.7   male        71
## 100      98.4 female        84
## 101      98.5 female        83
## 102      98.6 female        82
## 103      98.6 female        85
```

Que peut-on tirer comme conclusions rigoureuses à partir de ces données ?

- Quelle est la valeur moyenne du rythme cardiaque ?
- Quelle confiance accorder à cette estimation ?
- Le rythme cardiaque est-il différent chez les hommes et les femmes ?
- Le rythme cardiaque dépend-il de la température corporelle ? ...

Comment répondre à ces questions ? Graphiques ? Calculs ?

Un exemple: température corporelle

Le jeu de données **Heart Rate** renseigne pour 130 individus la température corporelle, le sexe and son rythme cardiaque.

```
##     BodyTemp Gender HeartRate
## 1       96.3   male        70
## 2       96.7   male        71
## 100     98.4 female        84
## 101     98.5 female        83
## 102     98.6 female        82
## 103     98.6 female        85
```

La statistique : étude d'observations répétées d'un phénomène aléatoire.

L'échantillonnage

- Une observation est associée à une variable mesurée sur un “individu statistique", lui-même membre d'un ensemble plus vaste appelé population, ou univers.
- En général on ne peut pas observer tout l'univers.
- On procède donc à un échantillonnage de la population : i.e. on observe sur un nombre limité d'individus prélevés au hasard au sein de la population, la valeur de la variable étudiée.
- On a alors recours à la théorie des probabilités pour formaliser mathématiquement l'échantillonnage aléatoire.

L'échantillonnage aléatoire

- Un phénomène dont les valeurs dépendent du hasard peut être modélisé par une variable aléatoire $X : \Omega \mapsto \mathbb{R}$, où Ω est l'univers abstrait et $\mathbb{R}$ (ou $\mathbb{R}^p$ ou un espace plus complexe) l'espace des réalisations.
  - si la loi de X est connue on peut alors calculer les probabilités des événements qui nous intéressent. Exemple ?
  - si la loi de X est inconnue ou non entièrement connue et que l'on dispose cependant d'une suite d'observations, on est dans le domaine de **la statistique**. Exemple ?

- Lorsque les valeurs observées $x_1, \ldots, x_n$ sur les individus prélevés sont des réalisations de variables aléatoires (v.a.) mutuellement indépendantes $X_1, \ldots, X_n$ ayant toutes la même loi, la suite $X_1, \ldots, X_n$ est appelée échantillon aléatoire de taille n (ou n-échantillon).

Démarche statistique

La démarche statistique comporte généralement deux étapes :

- **Phase de modélisation :** on suppose que l'observation X est un objet aléatoire dont la loi est inconnue mais appartient cependant à une famille spécifiée de lois $(P_\theta)_{\theta \in \Theta}$ que l'on appelle un modèle statistique.
- **Inférence :** À partir de la connaissance du modèle $(P_\theta)_{\theta \in \Theta}$ et de l'observation X, on détermine un maximum d'information sur la loi réelle des observations, et en premier lieu sur les paramètres en jeu dans le modèle.

Outline

Statistique descriptive : moyennes et variances empiriques

- Supposons que l'on observe un échantillon $X_1, \dots, X_n$ de même loi celle d'une v.a. X.
- Pour décrire la loi de X, on peut construire de nombreuses statistiques:
  - La moyenne (empirique) : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ qui donne la tendance centrale des observations,
  - La variance (empirique) : $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}(X_i - \bar{X})^2$ qui mesure la dispersion des observations.
- Ces quantités peuvent être vues comme des quantités approchant les quantités correspondantes pour la distribution sous jacente (celle de X), si elles existent :
  - Espérance de la loi de X : $\mathbb{E}(X)$
  - Variance de la loi de X : $\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2$

Statistique descriptive : moyennes et variances empiriques

- Supposons que l'on observe un échantillon $X_1, \ldots, X_n$ de même loi celle d'une v.a. X.
- Pour décrire la loi de X, on peut construire de nombreuses statistiques:
  - La moyenne (empirique) : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ qui donne la tendance centrale des observations,
  - La variance (empirique) : $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}(X_i - \bar{X})^2$ qui mesure la dispersion des observations.
- Ces quantités peuvent être vues comme des quantités approchant les quantités correspondantes pour la distribution sous jacente (celle de X), si elles existent :
  - Espérance de la loi de X : $\mathbb{E}(X)$
  - Variance de la loi de X : $\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2$

Question: Une v.a. a-t-elle toujours une espérance ?

Statistique descriptive : moyennes et variances empiriques

- Supposons que l'on observe un échantillon $X_1, \ldots, X_n$ de même loi celle d'une v.a. X.
- Pour décrire la loi de X, on peut construire de nombreuses statistiques:
  - La moyenne (empirique) : $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ qui donne la tendance centrale des observations,
  - La variance (empirique) : $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1} (X_i - \bar{X})^2$ qui mesure la dispersion des observations.
- Ces quantités peuvent être vues comme des quantités approchant les quantités correspondantes pour la distribution sous jacente (celle de X), si elles existent :
  - Espérance de la loi de X : $\mathbb{E}(X)$
  - Variance de la loi de X : $\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2$
- On peut construire de nombreuses autres statistiques de la même façon pour décrire la loi de X.
- La **statistique descriptive** a pour objet la construction d'indicateurs, de tableaux, et de représentations graphiques permettant de résumer visuellement les observations de l'échantillon.

Statistique descriptive : USA Renewable Energy Technical Potential dataset

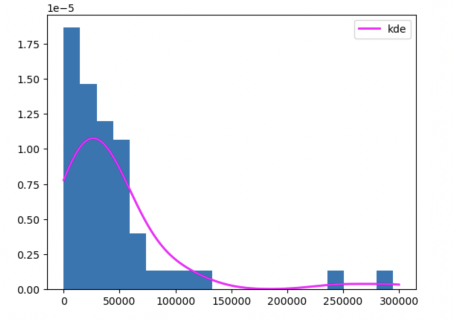
	urbanUtilityScalePV_GWh	urbanUtilityScalePV_GW	urbanUtilityScalePV_km2	ruralUtilityScalePV_GWh	rura
count	51.000000	51.000000	51.000000	5.100000e+01	
mean	43758.235294	23.431373	496.921569	5.502219e+06	
std	54365.369016	27.735360	577.748348	6.284523e+06	
min	8.000000	0.000000	0.000000	0.000000e+00	
25%	12162.000000	6.000000	134.000000	1.296446e+06	
50%	30492.000000	16.000000	338.000000	4.876185e+06	
75%	51824.000000	31.000000	659.500000	8.235158e+06	
max	294684.000000	154.000000	3213.000000	3.899358e+07	

Jeu de données étudié dans le TP StatExplo.

Statistique descriptive : distribution empirique

La distribution empirique de l'échantillon *approche* la distribution théorique de la loi des observations.

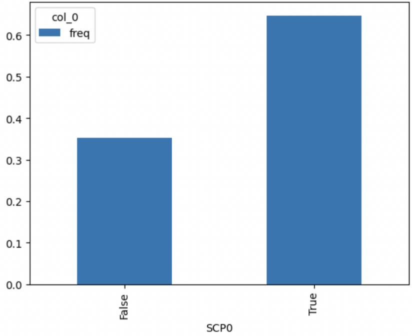
- Pour une distribution continue : histogramme des observations



Statistique descriptive : distribution empirique

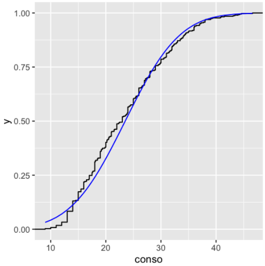
La distribution empirique de l'échantillon *approche* la distribution théorique de la loi des observations.

- Pour une variable discrète : diagrammes en bâtons des observations



Statistique descriptive : fonction de répartition empirique

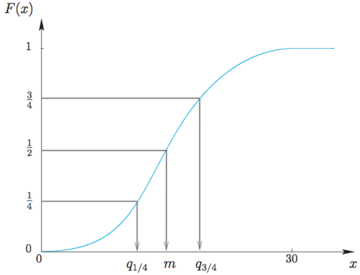
La fonction de répartition empirique de l'échantillon *approche* la fonction de répartie théorique de la loi des observations.



Question: Donner l'expression de la fonction de répartition empirique

Statistique descriptive : quantiles empiriques

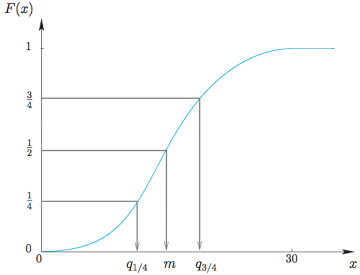
- La médiane m de la loi de la v.a. X est la valeur m telle que $P(X \geq m) = 1/2$.
- De façon plus générale, on appelle quantile d'ordre α de la la loi de X toute valeur q_α telle que $P(X \leq q_\alpha) = \alpha$, i.e. $F(q_\alpha) = \alpha$.



- Soient $x_1, \ldots, x_n$ des réalisations d'un échantillon aléatoire $X_1, \ldots, X_n$. Ces observations étant classées par ordre croissant $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, on appelle quantile empirique d'ordre α la $\lceil n\alpha \rceil$ -ème observation.

Statistique descriptive : quantiles empiriques

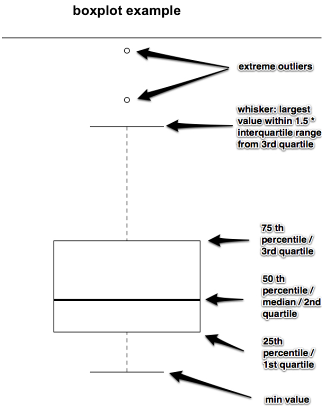
- La médiane m de la loi de la v.a. X est la valeur m telle que $P(X \geq m) = 1/2$.
- De façon plus générale, on appelle quantile d'ordre α de la la loi de X toute valeur q_α telle que $P(X \leq q_\alpha) = \alpha$, i.e. $F(q_\alpha) = \alpha$.



- Soient $x_1, \ldots, x_n$ des réalisations d'un échantillon aléatoire $X_1, \ldots, X_n$. Ces observations étant classées par ordre croissant $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, on appelle quantile empirique d'ordre α la $\lceil n\alpha \rceil$ -ème observation.

Statistique descriptive : boxplots

Un boxplot (boîte à moustache ou diagramme en boîte) est un résumé graphique de la distribution d'une variable :



Référence : blog Stochastic Nonsense

Outline

Expérience statistique

- Une **expérience statistique** est la donnée d'un objet aléatoire X à valeurs dans un espace mesurable muni d'une famille de loi $(P_\theta)_{\theta\in\Theta}$, supposée contenir la loi de la variable aléatoire X.
- Par exemple, X peut être une v.a. continue ou discrète.
- L'ensemble des lois $(P_\theta)_{\theta\in\Theta}$ est connu, mais **le paramètre** θ de la "vraie" loi est inconnu. La démarche statistique consiste à donner de l'information sur le paramètre θ en s'appuyant sur des observations du phénomène X.
- Modèle paramétrique : $\Theta \subset \mathbb{R}^d$ et θ est le paramètre du modèle.
- On appelle ***n*-échantillon** la donnée d'un n-uplet $X = (X_1, \ldots, X_n)$ de v.a. indépendantes et de même loi p_θ.

Examples

- **Activité / arrêt d'une éolienne.** On relève tous les jours le statut d'une élolienne (arrêt/activité) et on modélise le phénomène par une loi de Bernoulli de paramètre inconnu p.

  L'expérience statistique est modélisée par le modèle du pile ou face, i.e. des lois de Bernoulli $((\mu_p)^{\otimes n})_{p\in[0,1]}$ sur $\{0,1\}^n$.

- On mesure la **taille** de n individus et on modélise cette mesure par une loi gaussienne de paramètres inconnus.

  L'expérience statistique est modélisée par un modèle gaussien unidimensionnel sur $\mathbb{R}^n$ muni des lois gaussiennes $\left((\nu_{\mu,\sigma^2})^{\otimes n}\right)_{\mu\in\mathbb{R},\sigma>0}$ de moyenne μ et de variance σ^2.

- On mesure la **vitesse du vent** toutes les heures pendant une année sur un site donné et on modélise ces mesures par une distribution de Weibull de densité (sur $\mathbb{R}$) :

$$f(x;k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \mathrm{e}^{-(x/\lambda)^k}$$

  de paramètre de forme k et de paramètre d'echelle λ.

Outline

Estimateur

- On considère une expérience statistique X dans un modèle statistique paramétrique composé des lois $(P_\theta)_{\theta \in \Theta}$.
- Objectif : estimer une quantité $g(\theta)$ à partir de l'observation X, où g une application de Θ dans $\mathbb{R}^q$.
  Ex: $g(\mu, \sigma^2) = \mu$ dans le modèle gaussien précédent.
- Souvent on veut estimer θ lui-même et dans ce cas $g = Id$.

Definition

- *On appelle statistique toute variable (ou vecteur) aléatoire T à valeurs dans $\mathbb{R}^q$ qui est fonction de X :*

$$T = h(X)$$

  (où h est une application mesurable.)
- *Un estimateur $\hat{g}$ de $g(\theta)$ est une statistique à valeurs dans $g(\Theta)$, qui ne dépend pas de θ.*

Estimateurs : exemples

Dans le jeu du pile ou face :

- La quantité $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur de p qui semble pertinent (la loi des grands nombres s'applique...)
- La quantité $S_n^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ est un estimateur de la variance $p(1-p)$.

Estimateurs : exemples

Dans le jeu du pile ou face :

- La quantité $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur de p qui semble pertinent (la loi des grands nombres s'applique...)
- La quantité $S_n^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ est un estimateur de la variance $p(1-p)$.

mais :

- $\hat{p} = 0$ est aussi un estimateur (idiot) de p.

⚠ La définition d'un estimateur ne nous dit rien sur la qualité de l'estimation.

Fonction de vraisemblance (cas iid)

- On observe un n échantillon $X_1, \ldots, X_n$ sur $\mathbb{R}$ de loi commune p_θ de densité s_θ.
- La fonction de *vraisemblance* (*Likelihood* en anglais) pour le modèle paramétrique $(p_\theta)_{\theta \in \Theta}$ est la fonction définie sur Θ par

$$L_n(\theta) := \prod_{i=1}^{n} s_\theta(X_i)$$

- On considère aussi la fonction log-vraisemblance :

$$\ell_n : \theta \mapsto \ln(L_n(\theta)).$$

La méthode du maximum de vraisemblance (EMV)

- Méthode du maximum de vraisemblance : estimer le paramètre θ en choisissant le paramètre “le plus vraisemblable” pour l’échantillon disponible : i.e. trouver $\hat{\theta}_{\text{EMV}}$ pour lequel la vraisemblance $L_n(\theta)$ est maximale.
- On dit qu’un estimateur $\hat{\theta}$ est un estimateur du maximum de vraisemblance s’il vérifie

$$\hat{\theta} \in \operatorname{argmax}_{\theta \in \Theta} L_n(\theta).$$

EMV pour des lois classiques

- Pour la famille des lois **exponentielles**, la fonction de vraisemblance vaut $L_n(\lambda) = \lambda^n \prod_{i=1,\ldots,n} \exp(-\lambda X_i) \mathbb{1}_{\mathbb{R}^+}(X_i)$ et les X_i sont positifs p.s. On a $\ell_n(\lambda) = n \ln \lambda - \lambda \sum_{i=1,\ldots,n} X_i$ et $\hat{\lambda} = 1/\bar{X}_n$ est un EMV de λ.

- Pour un échantillon iid de loi **gaussienne** $\mathcal{N}(\mu, \sigma^2)$, on vérifie que $\bar{X}_n$ et $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ sont des EMV de μ et σ^2.

- Pour une loi **discrète**, la loi de référence à considérer est la mesure de comptage. Par ex pour une loi de **Poisson** $\mathcal{P}(\lambda)$, la densité par rapport à la mesure de comptage sur $\mathbb{N}$ vaut pour tout $k \in \mathbb{N}$:

$$f_\lambda(k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

  La log-vraisemblance pour un échantillon $X_1 \ldots, X_n$ de loi $\mathcal{P}(\lambda)$ vaut

$$\ell_n = -n(\lambda - \bar{X}_n \ln \lambda) - \sum_{i=1}^n \ln(X_i!),$$

  On vérifie alors que $\hat{\lambda} := \bar{X}_n$ dès que $\bar{X}_n > 0$.

EMV : un problème d'optimisation

$$\hat{\theta} \in \mathrm{argmax}_{\theta \in \Theta} L_n(\theta).$$

- L'EMV n'existe pas toujours.
- La solution de l'EMV n'est pas toujours explicite
- L'EMV n'est pas toujours unique.
- Mais pour des modèles "réguliers" , l'EMV est une approche "efficace" (dans un sens statistique à définir).
- Beaucoup de méthodes en Machine Learning reposent de cette méthode.

Application: estimation de distribution de la vitesse du vent

- On modélise des relevés de **vitesse de vent** par la distribution de Weibull de densité (sur $\mathbb{R}$) :

$$f(x; k, \lambda) = \frac{\beta}{\alpha} \left(\frac{x}{\alpha}\right)^{\beta-1} e^{-(x/\alpha)^k}$$

  de paramètre de forme β et de paramètre d'échelle α.

- On observe des vitesses $X_1, \ldots, X_n$ sur un même site sur plusieurs jours consécutifs.

- Pour simplifier on ne prend pas en compte la **dépendance** éventuelle entre les observations et on considère ces données comme un **échantillon i.i.d**.

Application: estimation de distribution de la vitesse du vent

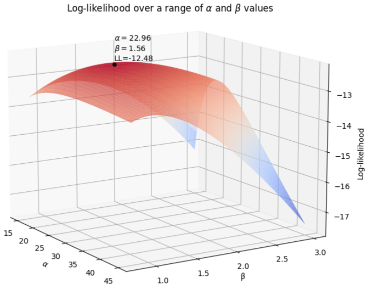
- On modélise des relevés de **vitesse de vent** par la distribution de Weibull de densité (sur $\mathbb{R}$) :

$$f(x; k, \lambda) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-(x/\alpha)^k}$$

de paramètre de forme β et de paramètre d'échelle α.

- On observe des vitesses $X_1, \ldots, X_n$ sur un même site sur plusieurs jours consécutifs.

- Estimation de la loi de Weibull par MLE
  - $\hat{\alpha} = \left(\frac{1}{n}\sum X_i^{\hat{\beta}}\right)^{\frac{1}{\hat{\beta}}}$
  - $\hat{\beta}$: solution numérique (méthode de Newton-Raphson)



Outline

Données Ozone

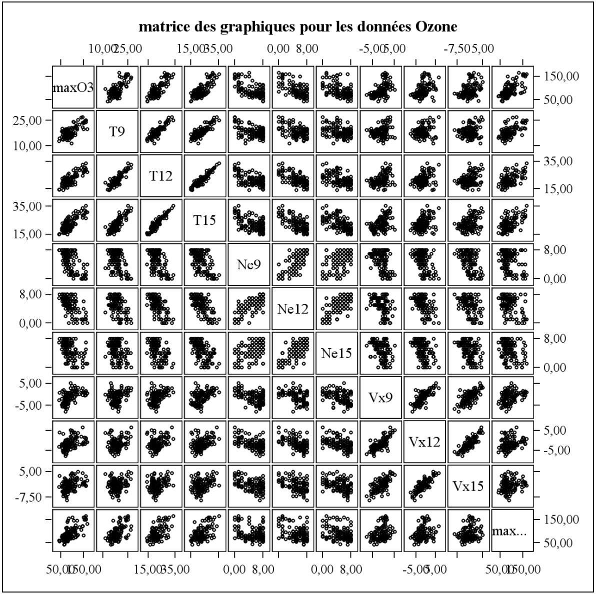
Jeu de données : niveaux de pollution enregistrés sur 112 journées de l'été 2001 à Rennes [1].

On souhaite expliquer la quantité d'ozone maxO3 par les mesures suivantes :

- T9, T12, T15 : température à 9h, 12h, 15h ;
- Ne9, N12, N15 : nébulosité à 9h, 12h, 15h ;
- Vx9, Vx12, VX15 : vitesse du vent sur un axe Est-Ouest à 9h, 12h, 15h ;
- MaxO3v : ozone mesurée la veille.

[1] exemple tiré de l'ouvrage *Statistiques avec R, P.A. Cornillon et al., PUR 2008*

Matrice du nuage de ces variables



Régression linéaire multiple

Dans le modèle de **régression multiple**, on suppose que l'espérance de Y est une fonction linéaire des variables explicatives x_j :

$$Y_i = \mathbb{E} Y_i + \varepsilon_i = \theta_0 + \theta_1 x_{i,1} + \cdots + \theta_{p-1} x_{i,p-1} + \varepsilon_i, \quad i = 1 \ldots n$$

où

- Y est la variable dite *dépendante* (à expliquer),
- $x_1, \ldots, x_{p-1}$ sont les variables explicatives ou regresseurs, supposées déterministes (design fixe)
  Remarque : la i-ème observation est le vecteur $(1, x_{i,1}, \ldots, x_{i,p-1})'$
- ε est le vecteur des erreurs, ces erreurs sont supposées indépendantes et de même loi centrée de variance σ^2.
- $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{p-1})' \in \mathbb{R}^p$ et $\sigma > 0$ sont les paramètres du modèle. Lorsque $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ on parle de régression linéaire multiple gaussienne.

Écriture matricielle du modèle de régression linéaire multiple

Il est possible d'écrire ce modèle sous la forme matricielle suivante :

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p-1} \end{bmatrix}}_{\boldsymbol{X}=[\boldsymbol{e},x_1,\dots,x_{p-1}]} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{p-1} \end{bmatrix}}_{\boldsymbol{\theta}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}}$$

$$Y = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

Modélisation de Y via une représentation "simple" de son espérance.

Problème des moindres carrés

Pour choisir θ on minimise la distance euclidienne entre Y et $\boldsymbol{X}\theta$:

Definition

- *On appelle estimateur des moindres carrés de* $\boldsymbol{X}\theta$ *un estimateur* $\widehat{\boldsymbol{X}\theta}$ *de* $\boldsymbol{X}\theta$ *qui minimise* $\|Y - \widehat{\boldsymbol{X}\theta}\|^2$.
- *On appelle estimateur des moindres carrés de* θ *un estimateur* $\hat{\theta}$ *qui minimise* $\|Y - \boldsymbol{X}\hat{\theta}\|^2$.

Solution des moindres carrés

Proposition

- *L'estimateur des moindres carrés de $\boldsymbol{X}\theta$ est unique et vérifie*

$$\widehat{\boldsymbol{X}\theta}_{MC} = Argmin\{\|Y - u\|^2 \mid u \in Im(\boldsymbol{X})\} = \mathbf{P}_{Im(\boldsymbol{X})} Y.$$

- *θ minimise $\|Y - \boldsymbol{X}\theta\|^2$ si et seulement si θ est solution de*

$$\boldsymbol{X}'\boldsymbol{X}\,\theta = \boldsymbol{X}'\,Y. \tag{1}$$

Les équations (1) sont appelées **équations normales** du modèle linéaire.

Cas régulier

Lemma

Si $\boldsymbol{X}$ est injective (cas régulier) alors $\boldsymbol{X}'\boldsymbol{X}$ est inversible et il existe une unique solution aux équations normales :

$$\hat{\theta}_{MC} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y.$$

On peut aussi vérifier que la matrice de projection sur l'image de $\boldsymbol{X}$ (aussi appelé *hat matrix* ou *hat operator*) vérifie :

$$\mathbf{P}_{\mathrm{Im}(\boldsymbol{X})} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

et on a bien que

$$\boldsymbol{X}\hat{\theta}_{\mathrm{MC}} = \widehat{\boldsymbol{X}\theta}_{\mathrm{MC}}.$$

Vecteur des valeurs prédites et MSE

- On définit le vecteur des valeurs prédites par (cas régulier)

$$\hat{Y} := \boldsymbol{X}\widehat{\boldsymbol{\theta}}.$$

Cette quantité $\hat{Y}$ peut être vue comme un estimateur de $\mathbb{E}Y$, ou comme un prédicteur du vecteur Y.

- Dans le cas régulier, au point de design x, le prédicteur vaut

$$\hat{y}(x) = x'\widehat{\boldsymbol{\theta}}.$$

- Application jeu de donnés ozone : si on dispose pour une nouvelle journée des données T9, T12, ... Max03v, on peut alors proposer une prédiction pour le maximum d'ozone sous la forme :

$$\hat{O}_3 = \hat{\theta}_0 + \hat{\theta}_1 T_9 + \hat{\theta}_1 T_{12} + \cdots + \hat{\theta}_{11} Max0_3 v$$

Vecteur des valeurs prédites et MSE

- On définit le vecteur des valeurs prédites par (cas régulier)

$$\hat{Y} := \boldsymbol{X}\widehat{\boldsymbol{\theta}}.$$

  Cette quantité $\hat{Y}$ peut être vue comme un estimateur de $\mathbb{E}Y$, ou comme un prédicteur du vecteur Y.

- Dans le cas régulier, au point de design x, le prédicteur vaut

$$\hat{y}(x) = x'\widehat{\boldsymbol{\theta}}.$$

- L'erreur moyenne quadratique (Mean Squared Error) vaut

$$MSE := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

  Cette quantité est égale au risque empirique pour la perte ℓ_2 du prédicteur linéaire $\hat{y}(x) = x'\hat{\theta}$.

Ce qu'il faut retenir de la régression linéaire

- Méthode standard et populaire pour la régression.
- Fournit un prédicteur de type linéaire : $\hat{y}(x) = x'\hat{\theta}$.
- Contrairement à de nombreuses méthodes plus complexes en apprentissage statistique, une description statistique fine de la régression linéaire est possible (non présentée dans cet exposé).
- Extension à la classification : régression logistique.
- Extension pour la grande dimension : le Lasso et ses variantes.