Statistiques avancées — Régression
Cours 1 : Rappels et régression linéaire simple

20 Septembre 2021

1 Présentation de l'option
- Organisation
- Matériel

2 Science des Données

3 Paradigme statistique

4 Exemples élémentaires

5 Régression linéaire simple

Programme de l'option

8 semaines de cours

- Statistiques avancées et régression
- Introduction à l'apprentissage supervisé
- Optimisation pour l'apprentissage
- Apprentissage non supervisé
- Introduction au Deep learning
- Data camp
- Introduction au traitement du langage naturel
- Analyse de données pour les processus industriels

Programme de l'option

8 semaines de cours

- Statistiques avancées et régression
- Introduction à l'apprentissage supervisé
- Optimisation pour l'apprentissage
- Apprentissage non supervisé
- Introduction au Deep learning
- Data camp
- Introduction au traitement du langage naturel
- Analyse de données pour les processus industriels

Organisation : équipe enseignante

Cours / PC

- Geneviève Robin CR CNRS, Université d'Évry Val d'Essonne, École Polytechnique (Cours)
- Mehdi Abou El Qassime, doctorant UM6P (PC et TP Python)

Présentation du cours

- Jour 1 : Introduction et rappel
  - Introduction générale à la science des données
  - Rappels de statistiques
  - Mise en place de l'environnement de travail
- Jour 2 : Régression linéaire multivariée
  - Estimateur du maximum de vraisemblance
  - Tests d'hypothèse et sélection de modèle
  - Régression linéaire robuste et régression ridge
- Jour 3 : Modèles linéaires généralisés
  - Famille exponentielle : Définition et exemples
  - Modèles linéaires généralisés (GLM) : Modèle et estimation
- Jour 4 : Modèles linéaires en grande dimension
  - Fléau de la dimension, non unicité du MLE et erreur d'estimation en grande dimension
  - L'estimateur du Lasso : définition, paramètre de régularisation et analyse théorique
- Jour 5 : Synthèse du cours et projet final
  - Exercices théoriques
  - Analyse d'un jeu de données

Organisation du cours

- **Cours :**
  - Présentation des méthodes statistiques
  - Exemples pratiques avec Python
  - Exercices théoriques
  - Exercices d'entraînement à la maison
- **Notation :**
  - DM (en groupe de 2-3 étudiants) : partie théorique, résumé d'un article et analyse d'un jeu de données

1 Présentation de l'option

2 Science des Données
- Caractéristiques des données modernes
- Statistiques et Machine learning
- Exemples

3 Paradigme statistique

4 Exemples élémentaires

5 Régression linéaire simple

Données modernes : explosion du volume disponible

$1 \text{ ZB} = 10^{12} \text{ GB}$

Données modernes : caractéristiques

- Dans tous les secteurs industriels et académiques : santé, biologie, finance, marketing, numérique, etc.
- Bases de données de grande dimension
  - Génomique : 10^3 patients, 10^6 gènes
  - Réseaux sociaux : 50×10^6 utilisateurs, 250×10^6 connexions
  - Écologie : 800 sites $\times$ 30 années $\times$ 30 espèces
- Grande hétérogénéité des données : tableaux, texte, images, réseaux, etc.
- Jeux de données publics
  - Données publiques françaises : `data.gouv.fr`
  - Kaggle : `kaggle.com/datasets`
  - UC Irvine Machine Learning repository : `uci.edu/ml/`

Données modernes : enjeux

- Scientifiques : génomique, cosmologie, etc.
- Sociétaux : équité des algorithmes, protection des données

Search query	Work experience	Education experience	Profile views	Candidate	Xing ranking
Brand Strategist	146	57	12992	male	1
Brand Strategist	327	0	4715	female	2
Brand Strategist	502	74	6978	male	3
Brand Strategist	444	56	1504	female	4
Brand Strategist	139	25	63	male	5
Brand Strategist	110	65	3479	female	6
Brand Strategist	12	73	846	male	7
Brand Strategist	99	41	3019	male	8
Brand Strategist	42	51	1359	female	9
Brand Strategist	220	102	17186	female	10

- Écologiques : hardware recyclable, consommation énergétique des data centers, etc.

Données modernes : exemple

- Science citoyenne pour la surveillance de la biodiversité :
  - Amateurs se greffent à des protocoles scientifiques
  - Comptent la faune/flore
  - Renseignent leurs comptes sur les plateformes dédiées
  - Organisations : Ligue pour la Protection des Oiseaux, International Waterbird Census (IWC), etc.
- "Big biodiversity data"
  - IWC : 50 ans de suivi, 25 000 sites écologiques, 143 pays, >150 espèces
  - Extraction de contenu Internet : Profusion d'informations géographiques, météorologiques, économiques

International Waterbird Census

Sitecodes	Sitenames	Pays	1990	1991	1992	1993	1994
Site 1	DZ00002	ALGERIE	42	28	57	33	130
Site 2	DZ00003	ALGERIE	131	#NA	#NA	33	116
Site 3	DZ00004	ALGERIE	126	131	91	263	320
Site 4	DZ00005	ALGERIE	292	130	182	95	54
Site 5	DZ00006	ALGERIE	#NA	#NA	214	648	170
Site 6	DZ00007	ALGERIE	#NA	#NA	#NA	#NA	22

Sites	Sitecode	Pays	Latitude	Longitude	Altitude_mean (m)	Dist_towns (m)
Site 1	DZ00002	ALGERIE	36.8310239001235	3.67681153688907	12,23	8121,330
Site 2	DZ00003	ALGERIE	33.493796409214	5.99047662601625	40,08	2345,407
Site 3	DZ00004	ALGERIE	36.53317578125	3.86765917968749	311,78	3305,642
Site 4	DZ00005	ALGERIE	36.3513057773426	2.56227090238024	289,28	10237,276
Site 5	DZ00006	ALGERIE	36.850379032258	3.72839264112904	58,90	3950,196
Site 6	DZ00007	ALGERIE	36.1479098800996	5.16352812986608	849,96	17563,138

Objectif : estimer la taille des populations pour informer les agences de conservation internationales

Méthodologie Statistique

- **Modèle statistique : une distribution sous-jacente à partir de laquelle les données sont tirées**
- Exemple : jeu de données de tailles de femmes Américaines
  - Jeu de données : ensemble de nombres indiquant les tailles
  - Statisticien : données viennent d'une distribution Gaussienne
  - Moyenne et écart-type caractérisent la distribution : le modèle

Machine Learning

- **Machine learning : les données sont utilisées pour entraîner un algorithme à faire une tache**
- Exemple : classification supervisée

Data Science

- Stockage et interrogation des données : expertise des bases de données
- Visualisation des données (Gephi, Tulip, widget python, etc.)
- **Modélisation statistique**
- Algorithms implementation : Python, R, H2O, TensorFlow, github, ...

Pour s'entraîner aux métiers de la Science des Données :

- `kaggle.com`, `datascience.net`
- notebooks python
- Coursera

Sciences de la vie

- modèles de durée de vie,
- modèles épidémiologiques,
- dynamique de population,
- génomique, protéomique.

Figure 2. Kaplan–Meier Estimates of Progression-free Survival.
The median duration of progression-free survival (indicated by the dotted lines) was 10.6 months in the group given irinotecan, fluorouracil, and leucovorin (IFL) plus bevacizumab, as compared with 6.2 months in the group given IFL plus placebo, corresponding to a hazard ratio for progression of 0.54 ($P<0.001$).

Géophysique

- Climat,
- Hydrologie,
- Energies fossiles

Contrary to much recent discussion, the latest corrected analysis shows that the rate of global warming has continued, and there has been no slow down.

Environnement

- pollution,
- météorologie,
- vent, ensoleillement...

Apprentissage statistique & traitement du signal

- Reconnaissance de parole
- Vision par ordinateur
- Traduction automatique

1 Présentation de l'option

2 Science des Données

3 Paradigme statistique
- Données
- Modélisation statistique
- Probabilités/statistiques

4 Exemples élémentaires

5 Régression linéaire simple

Problématique statistique

Point de départ : des données

$$(x_1, \ldots, x_n).$$

- numériques (scalaires ou vectorielles) : $x_i \in \mathbb{R}^d$
- symboliques (labels, catégorielles) : $x_i \in \{0, 1\}$ (par exemple)
- mixtes
- mais aussi des graphes, des fonctions, des textes, etc...

(a)

(b)

Problématique statistique

- Modélisation statistique :
  - Les données sont la réalisation d'un vecteur aléatoire

$$Z = (X_1, \ldots, X_n) \ , \quad Z(\omega) = (X_1(\omega), \ldots, X_n(\omega)) = (x_1, \ldots, x_n) \ .$$

  - Idée : Prendre en compte la variabilité des données analysées
    - bruit de mesure
    - échantillonnage
    - variabilité individuelle
    - ...
- Vocabulaire :
  - ω est un tirage/événement
  - $X_i(\omega)$ est une variable aléatoire

Modèle statistique : qu'est ce que c'est ?

- ON SUPPOSE que la loi du vecteur aléatoire

$$Z = (X_1, \ldots, X_n)$$

  est partiellement connue : cela reflète notre connaissance a priori du phénomène étudié.
- On traduit cette connaissance partielle en supposant que la loi inconnue de Z appartient à une famille $\mathcal{C}$ de lois, appelée modèle statistique
- Le modèle est dit paramétrique si $\mathcal{C} = \{\mathbb{P}_\theta \ : \ \theta \in \Theta\}$ où $\Theta \subset \mathbb{R}^q$.
- Le modèle est dit non-paramétrique si $\mathcal{C} = \{\mathbb{P}_\theta \ : \ \theta \in \Theta\}$ où $\Theta \subset \mathcal{H}$ est de dimension infini.
- Problématique : à partir d'une réalisation

$$x_1, \ldots, x_n$$

  et du modèle $\mathcal{C}$, on cherche à affiner notre connaissance de la loi de l'observation.

Modèle statistique

Définition (Modèle statistique)

Un modèle statistique est la donnée de :

- un espace mesurable $(\mathsf{Z}, \mathcal{Z})$, l'*espace des observations*,
- une famille de probabilités $\mathcal{C}$ sur $(\mathsf{Z}, \mathcal{Z})$.

Le modèle est dit paramétrique lorsque $\mathcal{C}$ correspond à une famille $\{\mathbb{P}_\theta\ ,\ \theta \in \Theta\}$, où Θ est un sous-ensemble de $\mathbb{R}^d$, avec $d \geq 1$; i.e., il existe une fonction associant à chaque $\theta \in \Theta \subset \mathbb{R}^d$ un élément $\mathbb{P}_\theta \in \mathcal{C}$.

Définition (Identifiabilité)

Le modèle est identifiable si la fonction $\theta \mapsto \mathbb{P}_\theta$ est injective, i.e. si $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ implique $\theta = \theta'$.

La définition précédente n'est pas vide de sens : dans la mesure où on observera uniquement des réalisations d'une loi $\mathbb{P}_\theta$, l'injectivité de $\theta \mapsto \mathbb{P}_\theta$ est fondamentale pour une inférence de θ.

Exemples :

- Le modèle Gaussien $\mathcal{N}(\mu, \sigma^2)$ avec $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$ est identifiable.
- Le modèle de pile ou face $\mathcal{B}(p)$ est identifiable.
- Le modèle linéaire : $Y = X\theta + \varepsilon$ où $Y \in \mathbb{R}^n$ $X \in \mathcal{M}(n, p)$ et $\varepsilon \sim \mathcal{N}(0, I_n)$ est identifiable si $X^t X$ inversible.

Statistique

Définition (Statistique)

Soient $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$ un modèle statistique et $(\mathsf{T}, \mathcal{T})$ un espace mesurable.
On appelle statistique sur $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$ une application mesurable T de $(\mathsf{Z}, \mathcal{Z})$ à valeurs dans $(\mathsf{T}, \mathcal{T})$.

Exemples : Si $Z = (X_1, \ldots, X_n)$ avec chaque $X_i \in \mathbb{R}^d$

- $S_1(Z) = \frac{X_1 + \ldots + X_n}{n}$
- $S_2(Z) = \min\{X_i^1, i = 1 \ldots n\}$
- $S_3(Z) = med(X_1, \ldots, X_n)$ pour $d = 1$

Définition (Statistiques indépendantes)

Nous dirons que les statistiques S et T sur $(\mathsf{Z}, \mathcal{Z}, \mathcal{C})$ sont indépendantes si pour toute loi $\mathbb{P} \in \mathcal{C}$, les éléments aléatoires S et T sont indépendants.

Statistique et Probabilité

- Probabilité : les lois sont supposées connues... Etant donnée une loi de probabilité $\mathbb{P}$ sur un espace mesurable $(\Omega, \mathcal{F})$ et un vecteur aléatoire Z, l'objet du calcul des probabilités est d'évaluer des quantités
  - de la forme
$$\mathbb{P}(f(Z) \geq c), \quad \mathbb{E}[f(Z)].$$
  - Temps moyen de retour d'une marche aléatoire à 0
  - Temps de convergence à l'équilibre d'un processus aléatoire.
  - ...
- Statistiques : on cherche à résoudre un problème inverse. Etant donné une **réalisation** $z = (x_1, \ldots, x_n)$ de l'observation $Z = (X_1, \ldots, X_n)$, on cherche à inférer
  - certaines caractéristiques de la loi de ce vecteur aléatoire.
  - tester des hypothèses statistiques.
  - ...

Les grands problèmes

A partir d'une réalisation de l'observation et d'un modèle statistique

- estimer : donner une valeur approchée d'une fonction de la loi (par exemple : sa moyenne, sa variance, mais nous serons amenés à évaluer des quantités bien plus sophistiquées) et donner une évaluation de l'erreur d'estimation (régions de confiance).
- tester une hypothèse sur la loi (exemple : $\mathcal{C} = \mathcal{C}_0 \cup \mathcal{C}_1$ peut-on dire que $\mathbb{P} \in \mathcal{C}_0$?)
- prédire une valeur encore inconnue et donner une évaluation de l'erreur de prédiction...

1 Présentation de l'option

2 Science des Données

3 Paradigme statistique

4 Exemples élémentaires
- Estimation d'une moyenne
- Sondage

5 Régression linéaire simple

Estimation d'une moyenne

Le problème le plus simple en statistique : on considère une densité de probabilité p (par rapport à λ_{Leb}) définie sur $\Theta = \mathbb{R}^d$ et centrée :

$$\int_{\mathbb{R}^d} xp(x)\mathrm{d}x = 0.$$

On considère ensuite le modèle de translation de paramètre θ dont la loi a pour densité par rapport à la mesure de Lebesgue :

$$p_\theta(x) = p(x - \theta).$$

L'expérience statistique induite par une observation X est

$$(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \{p_\theta \cdot \lambda_{\mathrm{Leb}} \ : \ \theta \in \Theta\})$$

Le n-échantillon $Z = (X_1, \ldots, X_n)$ correspond au modèle statistique

$$(\{\mathbb{R}^d\}^n, \mathcal{B}(\{\mathbb{R}^d\}^n), \{p_\theta^{\otimes n} \cdot \lambda_{\mathrm{Leb}}^{\otimes n} \ : \ \theta \in \Theta\})$$

Une application mesure légitime :

$$S(Z) = \frac{X_1 + \ldots + X_n}{n}.$$

Sondage

Il s'agit ici vraisemblablement d'un des cas les plus simples de statistique...

- population de N individus qui doivent élire les candidats A ou B.
- $N\theta$ votent A... la proportion $\theta \in \Theta = \{0, 1/N, \ldots, N/N\}$ est inconnue.
- Typiquement, N est très grand... on pratique donc un sondage : tirage sans remise de $n \ll N$ individus dans cette population.

Construction du modèle statistique

- Population de taille N.
- Échantillon (sans remise) de taille n.
- Espace de probabilité : Z $= \{0,1\}^n$ muni de la tribu des parties $\mathcal{Z} = \mathcal{P}(\{0,1\}^n)$.
- Observation : Statistique $Z = (X_1, \ldots, X_n)$ où pour tout $z = (x_1, \ldots, x_n)$, $X_i(z) = x_i$.
- Données : $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$ une réalisation de l'observation $Z = (X_1, \ldots, X_n)$:
  1. $x_i = 1$: le i-ème sondé vote A,
  2. $x_i = 0$: le i-ème sondé vote B.

Sondage

la Loi de l'observation : dépend du paramètre inconnu θ.

- Population totale N, $N\theta$ votent A, $(N - N\theta)$ votent B

$$\mathbb{P}_\theta(X_1 = x_1) = (N\theta)^{x_1}(N - N\theta)^{1-x_1}/N \quad x_1 \in \{0, 1\},$$

- Population totale $N - 1$, $N\theta - x_1$ votent A, $(N - 1 - (N\theta - x_1))$ votent B

$$\mathbb{P}_\theta(X_2 = x_2 | X_1 = x_1) = \frac{(N\theta - x_1)^{x_2}(N - 1 - (N\theta - x_1))^{1-x_2}}{N - 1},$$

Sondage

Au $(n+1)$-ème sondage, population totale $N-n$, $N\theta - \sum_{i=1}^{n-1} x_i$ votent A, $(N\theta - \sum_{i=1}^{n-1} x_i)$ votent B

$$\mathbb{P}_\theta(X_n = x_n | X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) = \frac{\left(N\theta - \sum_{i=1}^{n-1} x_i\right)^{x_n} \left(N - n - \left(N\theta - \sum_{i=1}^{n-1} x_i\right)\right)^{1-x_n}}{N-n}.$$

Sondage

Par conditionnement successif, la loi de l'observation $(X_1, \ldots, X_n)$ est donnée, pour tout $(x_1, \ldots, x_n) \in \mathsf{Z}$ par

$$\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}_\theta(X_1 = x_1)\mathbb{P}_\theta\left(X_2 = x_2 \,|\, X_1 = x_1\right) \times \mathbb{P}_\theta\left(X_n = x_n \,|\, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}\right) .$$

La loi de l'observation dépend d'un paramètre inconnu $\theta \in \Theta$ la proportion qui vote A.

Modèle statistique pour le sondage

- $\Omega = \{0,1\}^n$, l'espace des observations.
- $\mathcal{F} = \mathcal{P}(\{0,1\}^n)$, la tribu des observations (ici, l'ensemble des parties).
- Observation : $(X_1, \ldots, X_n)$ les variables canoniques

$$z = (x_1, \ldots, x_n) \in \{0,1\}^n \ , \quad X_i(z) = x_i \ .$$

- La mesure de comptage μ sur $\{0,1\}^n$ est définie, pour tout $A \in \mathcal{P}(\{0,1\}^n)$ par $\mu(A) = |(|\, A)$.
- La loi des observations a une densité par rapport à la mesure de comptage par

$$p(\theta, x_1, \ldots, x_n) = (N\theta)^{x_1}(N - N\theta)^{1-x_1}/N \times \ldots$$
$$\times \frac{\left(N\theta - \sum_{i=1}^{n-1} x_i\right)^{x_n}\left(N - n - \left(N\theta - \sum_{i=1}^{n-1} x_i\right)\right)^{1-x_n}}{N-n}.$$

Estimation de la proportion

- Pour estimer la proportion θ, il est naturel de "compter" le nombre $\hat{\theta}_n$ d'individus votant A dans l'échantillon

$$\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{1\}}(X_i) \ .$$

- Une telle statistique des observations est appelée un estimateur
- Le modèle statistique induit par $n\hat{\theta}_n$ est donné, pour tout $\theta \in \Theta = \{0, 1/N, \ldots, N/N\}$, par $\mathsf{Z} = \{0, \ldots, n\}$, $\mathsf{Z} = \mathcal{P}(\{0, \ldots, n\})$ et pour tout $k \in \{0, \ldots, n\}$,

$$\mathbb{P}_\theta(\hat{\theta}_n = k/n) = \frac{\binom{\lfloor N\theta \rfloor}{k}\binom{N - \lfloor N\theta \rfloor}{n-k}}{\binom{N}{n}} \ .$$

Estimation de la proportion

- population $N = 50 \times 10^6$,
- échantillon $n = 1000$,
- $\theta = 0.52$.
- Tracés de $k \longrightarrow \frac{\binom{\lfloor N\theta \rfloor}{k}\binom{N-\lfloor N\theta \rfloor}{n-k}}{\binom{N}{n}}$

- $\hat{\theta}_n$ donné par :

$$\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{X_i=1\}} \ .$$

  est un estimateur ponctuel du paramètre θ.
- Questions naturelles :
  - Comment quantifier l'erreur que nous commettons en estimant θ par $\hat{\theta}_n$ (régions de confiance) ?
  - Existe-t-il de meilleurs estimateurs ?
  - Peut-on tester l'hypothèse $\theta > 1/2$?

1 Présentation de l'option

2 Science des Données

3 Paradigme statistique

4 Exemples élémentaires

5 Régression linéaire simple
- Modèle et exemple
- Estimation
- Tests d'hypothèse

Le principe de la régression linéaire simple est de modéliser la relation linéaire entre deux variables, un "prédicteur" X et une "réponse" Y. Le modèle s'écrit :

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

où β_0 (l'intercept) et β_1 (la pente) sont des paramètres inconnus, et ε est la variable de bruit.

On suppose avoir accès à un n-échantillon $((X_1, Y_1), \ldots, (X_n, Y_n))$ consitué des n couples (X_i, Y_i), $1 \leq i \leq n$. On va chercher estimer les paramètres β_0, β_1 et σ^2 à partir de l'échantillon.

Exemple : Pollution de l'air

La pollution de l'air est un enjeu majeur en santé publique : le dioxyde de souffre (SO_2), le dioxyde d'azote (NO_2) et l'ozone (O_3) et la température sont surveillés quotidiennement dans l'air.

Nous souhaitons analyser ici la **relation linéaire** entre le maximum journalier de la concentration en ozone (en $\mu g/m^3$) et la température. Nous disposons de 112 données relevées durant l'été 2001 à Rennes.

Ici, le "prédicteur" X est la température journalière, et la réponse Y est le maximum journalier de la concentration en ozone. On dispose d'un 112-échantillon $((X_1, Y_1), \ldots, (X_{112}, Y_{112}))$. Le modèle est :

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

β_0 correspond à la moyenne de la concentration en ozone maximum lorsque la température est nulle (température de référence). β_1 correspond à l'"effet" de la température dans le modèle linéaire.

On s'intéresse à une méthode d'estimation pour les paramètres β_0 et β_1. L'estimateur le plus "simple" est l'estimateur dit des Moindres Carrés. Cet estimateur correspond à l'estimateur du maximum de vraisemblance dans le modèle Gaussien :

Définition (Estimateur des Moindres Carrés)

L'estimateur des moindres carrés $(\hat{\beta}_0, \hat{\beta}_1)$ est défini par :

$$(\hat{\beta}_0, \hat{\beta}_1) \in \mathrm{argmin}_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Calcul de l'estimateur des moindres carrés

On pose les quantités suivantes :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}), \qquad r = \frac{s_{XY}}{s_X s_Y}.$$

Les moindres carrés sont minimisés par :

$$\hat{\beta}_0 = \bar{Y} - \frac{s_{XY}}{s_X^2}\bar{X}, \quad \beta_1 = \frac{s_{XY}}{s_X^2}.$$

Ces estimateurs sont <u>sans biais</u> et de <u>variance minimale</u> parmi les estimateurs fonctions linéaires des Y_i. À chaque valeur de X_i correspond la valeur *estimée* de Y_i :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

Résidus et coefficient de détermination

On appelle *résidus* les quantités

$$E_i = Y_i - \hat{Y}_i,$$

pour $1 \leq i \leq n$. La variance inconnue σ^2 est estimée par la variation résiduelle

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} E_i^2.$$

On appelle *coefficient de détermination* la quantité

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$$

qui exprime le rapport entre la variance expliquée par le modèle et la variance totale.

Équivalence avec le maximum de vraisemblance

Sous l'hypothèse de bruit Gaussien, $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$. La fonction de vraisemblance de l'échantillon $(Y_1, \ldots, Y_n)$ s'écrit :

$$\mathcal{L}_n(\beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2\right].$$

On en déduit que la fonction de log-vraisemblance négative s'écrit :

$$\ell_n(\beta_0, \beta_1, \sigma^2) = \frac{n}{2}\log(\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Pour minimiser la log-vraisemblance négative, il faut minimiser la quantité $\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$, ce qui revient à calculer l'estimateur des moindres carrés.

On cherche à tester H_0 contre H_1 avec :

H_0 : $\beta_1 = 0$ (la température **n'a pas d'effet** sur la concentration en ozone) ;
H_1 : $\beta_1 \neq 0$ (la température **a un d'effet** sur la concentration en ozone).

Pour tester cette hypothèse, on doit trouver une statistique de test, ainsi que sa distribution sous l'hypothèse nulle $\beta_1 = 0$.

Lois des estimateurs

Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des variables aléatoires. Sous l'hypothèse des résidus Gaussiens, on peut montrer que les statistiques suivantes suivent respectivement des lois du χ^2 et de Student :

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2),$$

$$\frac{(\hat{\beta}_0 - \beta_0)}{s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}} \sim \text{Student}(n-2), \qquad \frac{(\hat{\beta}_1 - \beta_1)}{s\sqrt{\frac{1}{(n-1)s_X^2}}} \sim \text{Student}(n-2).$$

Ceci permet de tester l'hypothèse nulle $\beta_1 = 0$ (<u>ou</u> $\beta_0 = 0$, <u>séparément</u>), et de construire des intervalles de confiance.
En particulier, sous l'hypothèse nulle $\beta_1 = 0$, la statisti que de test $(n-2)\frac{R^2}{1-R^2}$ suit une distribution de Fisher $\mathcal{F}_{1,(n-2)}$.

Tests d'hypothèse

En utilisant le résultat précédent, et pour un niveau de significativité fixé α (niveau du test), le test suivant est de niveau α pour les hypothèses $H_0 \ : \ \beta_1 = 0$ contre $H_1 \ : \ \beta_1 \neq 0$:
On rejète H_0 si $F = \frac{R^2}{1-R^2} > f_{1;n-2;1-\alpha/2}$, où $f_{1;n-2;1-\alpha/2}$ est le $1-\alpha/2$ quantile de la distribution de Fisher à $(1, n-2)$ degrés de liberté.