Analyse, Classification et Indexation des données (ACID)

Régression linéaire simple et multiple

Akka Zemmari

LaBRI, Université de Bordeaux

2021 - 2022

Introduction

- Un des objectifs principaux des statistiques est d'expliquer la variabilité que l'on observe dans les données
- La régression linéaire est un outil statistique très utilisé pour étudier la présence de liens entre une variable *dépendante* Y et une ou plusieurs variables *indépendantes* $X_1, X_2, \cdots, X_p$.
- un modèle de régression peut servir à répondre à un des trois objectifs suivants :
  - décrire une réalité,
  - confronter des hypothèses : données provenant d'études expérimentales contrôles,
  - prédire.

Intuition[1]

Dans une entreprise, on a relevé les salaires des 32 employés (mensuel en euros, noté `sal`), ainsi que certaines caractéristiques socio-démographiques telles que l'ancienneté dans l'entreprise (en années, notée `anc`), le nombre d'années d'études après le bac (noté `apbac`), le sexe (`1 = F /2 = M` , noté `sex`), le type d'emplois occupés (en 3 catégories codées de 1 à 3, noté `emp`).

num	anc	sal	sex	apbac	emp
1	7	1231	1	3	2
2	15	1550	1	3	2
...	...	...	...	...	...
31	12	1539	2	2	1
32	13	1587	2	2	2

On souhaite alors évaluer l'effet éventuel des caractéristiques socio-démographiques sur le salaire des employés.

[1] exemple tiré de `https://www.math.univ-toulouse.fr/~barthe/M1modlin/poly.pdf`

Régression linéaire simple

Exemples

- On dispose de données relatives au taux d'absentéisme et au nombres d'employés dans des entreprises.
  On veut vérifier l'affirmation : plus le nombre d'employé est grand, plus le taux d'absentéisme augmente ...
- On dispose des chiffres des dépenses en carte de crédit et des revenus de personnes, y a-t-il un lien entre ces chiffres ?

Voir le notebook

Régression linéaire simple

Formulation

- Données : un échantillon de n paires (x_i, y_i) indépendants et identiquement distribués (i.i.d.)
- On cherche un modèle permettant de prédire les valeurs de $Y = (y_i)_{0 \leq i \leq n}$ en fonction des valeurs de $X = (x_i)_{0 \leq i \leq n}$:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Hypothèses

- H1 : X et Y sont des grandeurs numériques mesurées sans erreur. X est une donnée (exogène) dans le modèle, Y est aléatoire par l'intermédiaire de ε (c.-à-d. la seule erreur que l'on a sur Y provient des insuffisances de X à expliquer ses valeurs dans le modèle).
- H2 : Hypothèses sur le terme aléatoire. Les ε_i sont i.i.d.
  - (H2.a) En moyenne les erreurs s'annulent, $\mathbb{E}(\varepsilon_i) = 0$.
  - (H2.b) La variance de l'erreur est constante et ne dépend pas de l'observation : homoscédasticité $\mathbb{V}ar(\varepsilon_i) = \sigma_\varepsilon^2$
  - (H2.c) En particulier, l'erreur est indépendante de la variable exogène $COV(x_i, \varepsilon_i) = 0$.
  - (H2.d) Indépendance des erreurs, les erreurs relatives à 2 observations sont indépendantes (elles ne sont pas corrélées).
  - (H2.e) $\varepsilon_i \sim \mathcal{N}(0, \varepsilon)$.

Régression linéaire simple

Formulation

- Modèle de régression simple :

$$y_i = \beta_1 * x_i + \beta_0 + \varepsilon_i, \ \forall i$$

ce que l'on peut écrire $Y = X\beta + \varepsilon$ avec :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \text{ et } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

Régression linéaire simple

Formulation

- Objectif : trouver des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ de β_0 et β_1 respectivement.
- $\rightarrow$ Trouver les valeurs de β_0 et β_1 qui minimisent les écarts entre les valeurs réelles et les valeurs prédites.
  Plus formellement, il s'agit de minimiser la fonction :

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2.$$

comment faire ?
- on peut utiliser une descente de gradient, comment ?
- on peut résoudre un système d'équations (voir le détail des calculs au tableau).

Régression linéaire simple

Formulation

- Objectif : trouver des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ de β_0 et β_1 respectivement.
  Tout calcul fait :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x_i})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \text{ et } \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

  avec

$$\overline{x} = \frac{1}{n} \sum_{i=0}^{n} x_i \text{ et } \overline{y} = \frac{1}{n} \sum_{i=0}^{n} y_i.$$

Graphiquement

- Voir le schéma au tableau.
- voir le `jupyter notebook`.

Analyse des sources de variabilités

Intuition

Objectif de la régression : minimiser

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

avec y_i : la vérité terrain et $\hat{y}_i$ la valeur prédite par le modèle (Voir le schéma et les explications au tableau).

A partir de quand peut-on dire que le modèle est de "bonne qualité" ?

Analyse des sources de variabilités

La variabilité des données se décompose en une partie expliquée par le modèle de régression et une autre résiduelle (on parle de terme d'erreur). Ce que l'on peut écrire :

$$SCT = SCR + SCE$$

- SCT : somme des carrés totaux (il s'agit de la variance)

$$SCT = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- SCE : somme des carrés expliqués par le modèle

$$SCE = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

- SCR : somme des carrés résiduels, non expliqués par le modèle

$$SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

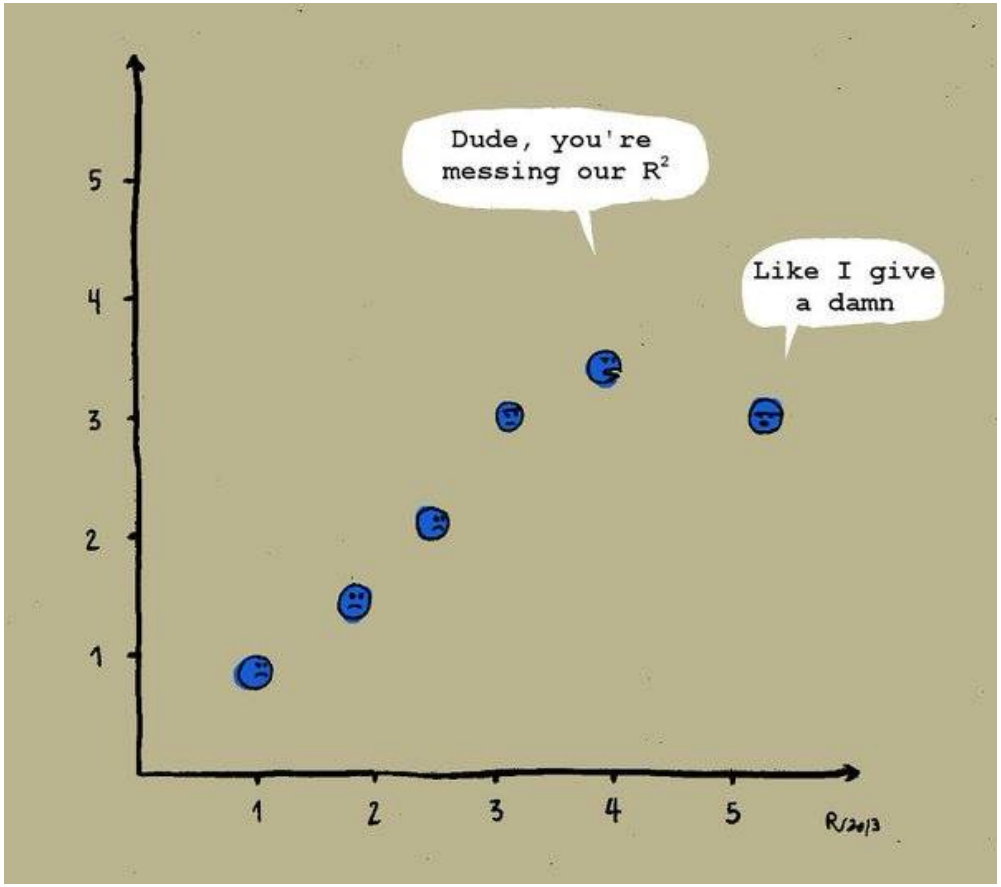
Coefficient de détermination

Le coefficient de détermination R^2 est défini par :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_{i=0}^{n} (y_i - \hat{y}_i)}{\sum_{i=0}^{n} (y_i - \overline{y})}$$

Ce coefficient mesure la variabilité expliquée par le modèle de régression linéaire.
Il vérifie toujours $0 \leq R^2 \leq 1$. En particulier :
- $R^2 \to 1$: le modèle est excellent
- $R^2 \to 0$: le modèle ne sert à rien.

Que mesure R^2 ?



Que mesure R^2 ?

Attention :

$$\text{"}R^2 \to 1 : \text{ le modèle est excellent"}$$

De quel modèle parle-t-on ?

Intuition : Quel est le R^2 d'un modèle entraîné sur un nuage composé de deux points ?

La volumétrie des données est-elle suffisante pour utiliser le modèle pour la prédiction ?

Test de significativité globale du modèle

On a notre modèle :

$$Y = \beta_1 * X + \beta_0$$

On teste alors l'hypothèse :

$$H_0 : \beta_1 = 0$$

contre l'hypothèse alternative

$$H_1 : \beta_1 \neq 0$$

Question : décrire ces hypothèses avec des "phrases".

Test de significativité globale du modèle

Tableau d'analyse de variance

Source de variation	Somme des carrés	DdL[2]	Carrés moyen
Régression (expliqués)	$SCE = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$	1	$SCE/1$
Résidus	$SCR = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$	$n-2$	$SCR/(n-2)$
Total	$SCT = \sum_{i=1}^{n} (y_i - \overline{y})^2$	$n-1$	

Statistique de test

$$F = \frac{SCE/1}{SCR/(n-2)} = \frac{R^2}{\frac{(1-R^2)}{(n-2)}} \equiv F(1, n-2)$$

Région critique au risque α

$$F > F_{1-\alpha}(1, n-2)$$

Il est temps de voir un exemple

Voir le tableau et le `jupyter notebook`.

Prévision et intervalle de prévision

Prévision ponctuelle

Pour un individu i^*, la prédiction ponctuelle s'écrit :

$$\hat{y_{i^*}} = \hat{y}\left(x_{i^*}\right)$$

L'erreur de prévision est alors

$$\hat{\varepsilon}_{i^*} = \hat{y}_{i^*} - y_{i^*}$$

avec $y_{i^*} = \beta_1 x_{i^*} + \beta_0 + \varepsilon_{i^*}$ et $\hat{y}_{i^*} = \beta_1 x_{i^*} + \beta_0 + \varepsilon_{i^*}$.
On démontre alors que

- $\mathbb{E}\left(\hat{\varepsilon}_{i^*}\right) = 0$ (exercice).
- $\mathbb{V}ar\left(\hat{\varepsilon}_{i^*}\right) = \sigma^2_{\hat{\varepsilon}_{i^*}}$ (voir Giraux & Chaix (1994)).

Prévision et intervalle de prévision

Prévision ponctuelle

Tout calcul fait :

$$\hat{\sigma}^2_{\hat{\varepsilon}_{i*}} = \hat{\sigma}^2_{\varepsilon}\left(1 + \frac{1}{n} + \frac{(x_{i*} - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right)$$

Prévision et intervalle de prévision

Prévision ponctuelle

La variance de l'erreur sera d'autant plus faible que :

1. $\hat{\sigma}_\varepsilon^2 = \frac{SCR}{n-2}$ est petit, i.e., la droite ajuste bien le nuage de points.
2. $(x_{i*} - \overline{x})^2$ est petit, i.e., le point est proche du centre de gravité du nuage.
3. $\sum_i (x_i - \overline{x})^2$ est grande, i.e., la dispersion des points est grande.
4. n est grand, i.e., le nombre d'observations ayant servi à construire le modèle est élevé.

Prévision et intervalle de prévision

Prévision par intervalle

On sait que $\varepsilon \sim \mathcal{N}(\prime, \sigma_\varepsilon)$.
On en déduit un intervalle de confiance au niveau $1-\alpha$ pour la prévision :

$$\hat{y}_{i^*} \pm t_{1-\alpha/2} \times \hat{\sigma}_{\hat{\varepsilon_{i^*}}}.$$

Régression linéaire multiple

- Généralisation de la régression linéaire simple au cas $p \geq 3$.
- On dispose donc d'une réalisation :

Observation	Y	X_1	X_2	$\cdots$	X_p
1	y_1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,p-1}$
2	y_2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,p-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,p-1}$

Régression linéaire multiple

- On cherche à établir un modèle linéaire :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

- Il faut alors estimer les paramètres $(\beta_i)_0 \leq i \leq p$ du modèle.
- Solution : Utiliser, là encore, la méthode des moindres carrés. Ce qui revient à minimiser la quantité :

$$\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)$$

Régression linéaire multiple

- On cherche à établir un modèle linéaire :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

- Une fois les $\hat{\beta}_i$ calculés, on détermine le coefficient de détermination

$$R^2 = \frac{SCR}{SCT}$$

Comme pour la régression simple, R^2 mesure la qualité du modèle.

Régression linéaire multiple

- On cherche à établir un modèle linéaire :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

- On cherche à décider si le modèle est pertinent ou pas. On pose alors le test de Fisher

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Contre l'hypothèse alternative

$$H_1 : \exists j \text{ tel que } \beta_j \neq 0.$$

Attention à l'interprétation (voir les explications).

Régression linéaire multiple

Comme pour la régression simple, on dresse le tableau (un peu modifié) de la variance :

Source de variation	Somme des carrés	DdL[3]	Carrés moyen
Régression (expliqués)	$SCE = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$	p	SCE/p
Résidus	$SCR = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$	$n-p+1$	$SCR/(n-p+1)$
Total	$SCT = \sum_{i=1}^{n} (y_i - \overline{y})^2$	$n-1$	

Et on calcule la statistique : $F_{obs} = CMR/CME$.
Au risque α, on lit la valeur critique $F_{1-\alpha,p,n-p+1}$ (dans une table de Fisher ou en utilisant un logiciel/instruction adéquate).

[3] Degrés de liberté

Régression linéaire multiple

Règle de décision :
Si $| F_{obs} | \geq F_{1-\alpha,p,n-p+1}$ alors on rejette H_0 au risque α.
Si on rejette H_0:

- c'est pas fini ... cela veut juste dire qu'il existe au moins un β_i qui est non nul ...
- $\Rightarrow$ Pour chacun des β_i, on doit tester l'hypothèse nulle :

$$H_0 : \beta_i = 0.$$

- on calcule pour cela la statistique : $t_{obs} = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$.
- On lit la valeur critique $t_{1-\frac{\alpha}{2},n-p+1}$ d'une loi de Student avec $n-p+1$ DdL.
- Si $| t_{obs} | \geq t_{1-\frac{\alpha}{2},n-p+1}$ alors on rejette H_0 au risque α.

Régression linéaire multiple

Dans la pratique (du moins dans ce cours) :

Il faut savoir :

1. Utiliser les bonnes instructions `Python` pour charger les données, faire la régression, récupérer les β_i,
2. générer le rapport de l'analyse , et (SURTOUT) SAVOIR LE LIRE :
   - y lire la significabilité du modèle,
   - poser les bonnes hypothèses à tester,
   - déduire le bon modèle à retenir (s'il en existe un).

Il est temps de voir un exemple

Voir le tableau et le `jupyter notebook`.