Présentation des modèles linéaires

C. Tuleau-Malot

Plan

Plan

Plan

Contexte d'étude

On s'intéresse à la modélisation et plus précisément aux modèles linéaires ou aux cas connexes.
On se limite au cadre des méthodes paramétriques où l'on considère des combinaisons de variables explicatives.
Cadre général :

- Une variable réponse Y quantitative
- p variables explicatives $X^1, \ldots, X^p$ quantitatives ou qualitatives

Méthodes :

- régression linéaire
- analyse de la variance et de la covariance

Contexte d'étude (2)

La théorie vous sera détaillée longuement dans de nombreux autres cours qui vont vous être dispensés cette année.
Deux approches : avec ou sans hypothèse de loi (généralement la normalité qu'il est assez difficile de vérifier).
D'autres hypothèses, adaptées aux méthodes appliquées, doivent elles être validés pour définir la qualité des résultats (homoscédasticité, colinéarité, etc)

Régression linéaire simple

Cette partie a pour unique objectif de revenir sur des concepts clés :

- modèle
- estimation
- test

Modèle

Soit Y la variable aléatoire à expliquer et X la variable explicative déterministe.
Le modèle revient à supposer que $\mathbb{E}[Y]$ est une fonction affine de X :

$$\mathbb{E}[Y] = f(X) = \beta_0 + \beta_1.X$$

Rq : Pour simplifier les choses, on a choisi de considérer X comme étant déterministe, mais on aurait pu le considérer comme aléatoire et alors on aurait eu que $\mathbb{E}[Y|X = x] = \beta_0 + \beta_1.x$.

Modèle (2)

Soit $\{(x_i, y_i),\ i \in \{1, \ldots, n\}\}$ n observations aléatoires et identiquement distribuées.
Le modèle s'écrit :

$$y_i = \beta_0 + \beta_1 . x_i + \varepsilon_i \quad i \in \{1, \ldots, n\}$$

ou encore

$$\mathbf{y} = \mathbb{X}\beta + \boldsymbol{\varepsilon}$$

Modèle (3)

Les hypothèses relatives à ce modèle :

- l'erreur est centrée et de variance constante
- β_0 et β_1 sont constantes
- hypothèse supplémentaire pour la partie inférence : $\varepsilon \sim \mathcal{N}(0, \sigma^2.\mathbb{I}_p)$

Estimation

Pour estimer les paramètres β_0, β_1 et σ^2, deux techniques :

- méthode des moindres carrés
- méthode du maximum de vraisemblance

Méthode des moindres carrés

critère :

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 . x_i)^2$$

Estimation (2)

Notations :

- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$
- $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- $s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$
- $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

La solution de la méthode des moindres carrés est :

$$\hat{y} = \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 . x$$

avec :

- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 . \bar{x}$
- $\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$

Estimation (3)

On montre que $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1.\bar{X}$ et $\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$ sont des estimateurs sans biais et de variance minimale parmi les estimateurs fonctions linéaires des y_i.

- prédictions : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1.x_i$
- résidus : $e_i = y_i - \hat{y}_i$

estimateur sans biais de la variance :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

Inférence

La matrice de covariance associée au couple de variables aléatoires $(\hat{\beta}_0, \hat{\beta}_1)$ est donnée par :

$$\Gamma = \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} & -\frac{\bar{x}}{(n-1)s_x^2} \\ -\frac{\bar{x}}{(n-1)s_x^2} & \frac{1}{(n-1)s_x^2} \end{pmatrix}$$

Sous l'hypothèse de normalité des résidus, on montre que

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

et donc

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)^{1/2}} \sim t_{n-2}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}\left(\frac{1}{(n-1)s_x^2}\right)^{1/2}} \sim t_{n-2}$$

Inférence

Intérêt :

- pouvoir tester la nullité des paramètres
- pouvoir donner des intervalles de confiances :
  - $\hat{\beta}_0 \pm t_{1-\alpha/2;n-2}\hat{\sigma}\left(\frac{1}{n}+\frac{\bar{x}^2}{(n-1)s_x^2}\right)^{1/2}$
  - $\hat{\beta}_1 \pm t_{1-\alpha/2;n-2}\hat{\sigma}\left(\frac{1}{(n-1)s_x^2}\right)^{1/2}$

Attention : Pour pouvoir obtenir une inférence conjointe sur $\hat{\beta}_0$ et $\hat{\beta}_1$, il faut travailler sur la loi du couple. L'intervalle de confiance associé est une ellipse d'équation :

$$n(\hat{\beta}_0-\beta_0)^2+2(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)\sum_{i=1}^{n}x_i+(\hat{\beta}_1-\beta_1)^2\sum_{i=1}^{n}x_i^2 = 2\hat{\sigma}^2\mathcal{F}_{1-\alpha;2;n-2}$$

Qualité d'ajustement

Notations :

- SST : $(n-1)s_y^2$ (total sum of squares)
- SSR : $(n-1)\frac{s_{xy}}{s_x^2}$ (regression sun of squares)
- SSE : $(n-2)\hat{\sigma}^2$ (error sum of squares)

On vérifie que SST=SSR+SSE
Coefficient de détermination :

$$R^2 = (r)^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{SSR}{SST}$$

où r est le coefficient de corrélation
Rq : Sous l'hypothèse $\beta_1 = 0$, la statistique $(n-2)\frac{R^2}{1-R^2} = (n-2)\frac{SSR}{SST}$ suit une loi de Fisher à 1 et $(n-2)$ degrés de liberté.

Prédiction

Soit x_0 une valeur pour les variables explicatives. Il existe deux façons de définir une intervalle de confiance de prédiction. Le premier est basé sur une encadrement de $\mathbb{E}[Y]$ sachant $X = x_0$ et le second est basé sur un encadrement de $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1.x_0$ qui tient compte de la variabilité individuelle.

- $\hat{y}_0 \pm t_{1-\alpha/2;n-2}\hat{\sigma}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2}\right)^{1/2}$
- $\hat{y}_0 \pm t_{1-\alpha/2;n-2}\hat{\sigma}\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2}\right)^{1/2}$

Un peu plus loin : régression non paramétrique

On considère un modèle de la forme :

$$y_i = f(x_i) + \varepsilon_i$$

avec une erreur centrée et une fonction f régulière.
Deux méthodes :

- Spline

$$\hat{f}_\lambda = \underset{f}{argmin} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_{-\infty}^{+\infty} (f''(x))^2 dx \right)$$

- Noyau

$$\hat{f}_\lambda(x) = \sum_{i=1}^{n} \frac{K(\frac{x-x_i}{\lambda})x_i}{\sum_{j=1}^{n} K(\frac{x-x_j}{\lambda})}$$

Influence

Le critère des moindres carrés est sensible à la présence de "outliers". Si une analyse statistique descriptive permet d'en détecter certains, il faut malgré tout procéder à une analyse plus fine et adapté au modèle linéaire.
En effet, il convient de repérer les observations (x_i, y_i) dites "influentes", à savoir celles pour lesquelles une petite variation génère une modification importante des caractéristique du modèle.

Influence des y_i

On peut écrire :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 . x_i = \sum_{j=1}^{n} h_{ij} y_j$$

avec $h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$.
Si l'on pose $\mathbb{H}$ la matrice des h_{ij}, on peut encore écrire :

$$\hat{\mathbf{y}} = \mathbb{H}\mathbf{y}$$

Les éléments diagonaux de $\mathbb{H}$ mesurent l'importance du rôle que joue y_i dans l'estimation de $\hat{y}_i$.
Autrement dit, l'écart de x_i à $\bar{x}$ joue un rôle important.

Les résidus

La définition des résidus n'est pas unique. Tout dépend des propriétés que l'on veut conférer aux résidus !

- *residus* : $e_i = y_i - \hat{y}_i$
- *residus*$_i$: $e_{(i)i} = y_i - \hat{y}_{(i)i} = \frac{e_i}{1-h_{ii}}$ avec $\hat{y}_{(i)i}$ la prévision de y_i déterminée sans l'observation (x_i, y_i).
  On définit :
  $$PRESS = \sum_{i=1}^{n} e_{(i)i}^2$$
- *residus standardise* : même en cas d'homoscédasticité, on a
  $$\mathbb{E}[e_i] = 0 \quad \text{et} \quad Var[e_i] = \sigma^2(1 - h_{ii})$$
  La version standardisé vise à avoir des résidus de même variance.
  $$r_i = \frac{e_i}{\sigma\sqrt{1 - h_{ii}}}$$

Les résidus (2)

- *residus studentise* : Dans la version standardisé, il y a un problème d'indépendance car σ dépend de e_i.
  Une estimation sans biais et indépendante de e_i est donnée par :

$$\sigma^2_{(i)} = \left[(n-2)\sigma^2 - \frac{e_i^2}{1-h_{ii}}\right] /(n-3)$$

  La version studentisée est :

$$t_i = \frac{e_i}{\sigma_{(i)}\sqrt{1-h_{ii}}}$$

  Rq : La version studentisé permet de détecter des observations "outliers" en comparant leur valeur à ± 2.

Influence

Un critère permettant d'évaluer l'influence d'une observation sur certain paramètre est :

- La Distance de Cook : $D_i = \frac{\sum_{j=1}^{n}(\hat{y}_{(i)j} - \hat{y}_j)^2}{2\sigma^2}$

Modèle

Soit $\{(x_i^{(1)}, \ldots, x_i^{(p)}, y_i),\ i \in \{1, \ldots, n\}\}$ n observations aléatoires et identiquement distribuées.
Le modèle s'écrit :

$$y_i = \beta_0 + \beta_1 . x_i^{(1)} + \ldots + \beta_p . x_i^{(p)} + \varepsilon_i \quad i \in \{1, \ldots, n\}$$

ou encore

$$\mathbf{y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

avec $\mathbb{X}$ une matrice à n lignes et $(p+1)$ colonnes

Moindres carrés

Le critère à minimiser est :

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i^{(1)} - \ldots - \beta_p x_i^{(p)})^2$$

soit encore :

$$\|\mathbf{y} - \mathbb{X}\boldsymbol{\beta}\|^2$$

Par dérivation matricielle, on obtient :

$$\mathbb{X}'\mathbf{y} - \mathbb{X}'\mathbb{X}\boldsymbol{\beta}$$

Sous l'hypothèse d'inversibilité de $\mathbb{X}'\mathbb{X}$ (autrement dit $\mathbb{X}$ de rang plein), on a :

$$\hat{\boldsymbol{\beta}} = \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbf{y}$$

Moindres carrés (2)

Ainsi on a :

$$\hat{\mathbf{y}} = \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'\mathbf{y} = \mathbb{H}\mathbf{y}$$

où $\mathbb{H} = \mathbb{X}\left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbb{X}'$: matrice de projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace vectoriel engendré par les colonnes de $\mathbb{X}$, noté $Vect(\mathbb{X})$.

On a :

$$e = \mathbf{y} - \hat{\mathbf{y}} = (\mathbb{I} - \mathbb{H})\mathbf{y}$$

projection orthogonale de $\mathbf{y}$ sur l'orthogonal de $Vect(\mathbb{X})$ dans $\mathbb{R}^n$.

Moindres carrés (3)

Propriétés :

- $\hat{\boldsymbol{\beta}}$: estimateur sans biais
- $\hat{\boldsymbol{\beta}}$: de matrice de variance-covariance $\sigma^2 \left(\mathbb{X}'\mathbb{X}\right)^{-1}$
- la matrice de variance-covariance de $\hat{\mathbf{y}}$ est $\sigma^2\mathbb{H}$
- la matrice de variance-covariance de e est $\sigma^2(\mathbb{I} - \mathbb{H})$
- estimateur sans biais de σ^2 : $\hat{\sigma}^2 = \frac{\|e\|^2}{n-p-1}$

Qualité

La qualité du modèle peut se quantifier à l'aide du coefficient de détermination qui a pour expression

$$R^2 = \frac{SSR}{SST}$$

avec

- $SST = \|\mathbf{y} - \bar{y}\mathbb{I}\|^2$
- $SSR = \|\hat{\mathbf{y}} - \bar{y}\mathbb{I}\|^2$
- $SSE = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \|e\|^2$

Rq : Ce coefficient de détermination peut être vu comme le cosinus au carré de l'angle entre $\mathbf{y}$ et $\hat{\mathbf{y}}$.

Inférence sur un coefficient

On a :

$$\frac{\hat{\beta}_i - \beta_i}{\sigma_{\hat{\beta}_i}} \sim T(n-p-1)$$

où $\sigma_{\hat{\beta}_i}$ est le i-ème terme diagonal de la matrice de corrélation de $\hat{\boldsymbol{\beta}}$.

Cette propriété permet de réaliser des tests ou un intervalle de confiance :

$$\hat{\beta}_i \pm t_{1-\alpha/2,(n-p-1)}\sigma_{\hat{\beta}_i}$$

Inférence sur le modèle

Dans ce contexte, on veut tester $\mathbb{H}_0 : \ \beta_1 = \beta_2 = \ldots = \beta_p = 0$.
La statistique de test qui intervient est

$$\frac{\|\hat{\mathbf{y}} - \bar{y}\mathbb{I}\|^2/p}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-p-1)} \sim \mathcal{F}(p; n-p-1)$$

Inférence sur un modèle réduit

Dans ce contexte, on veut tester la nullité simultanée de q coefficients ($q < p$).
Par simplicité, notre hypothèse nulle est
$\mathbb{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0$.
La statistique de test est alors :

$$\frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_q\|^2/q}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-p-1)} \sim \mathcal{F}(q; n-p-1)$$

Sélection de variables, choix de modèles

Il existe deux stratégies :

- Descriptive : identifier toutes les variables qui sont potentiellement explicatives
- Prédictive : le seul but est de minimiser une erreur. Cette stratégie intègre de la parcimonie

Plusieurs critères pour réaliser la sélection :

- R^2 ajusté
- statistique de Fisher
- C_p de Mallows

R^2 ajusté

Le R^2 est une fonction croissante du nombre de variables.

D'où la nécessité d'introduire une pénalisation par le nombre de variables :

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

statistique de Fisher

On considère la statistique partielle de Fisher définie par :

$$\frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_q\|^2/q}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-p-1)} = \frac{R^2 - R_q^2}{1 - R^2}\frac{n-p-1}{q}$$

De ce fait, si l'écart entre R^2 et R_q^2 est significatif, l'ajout des q variables est justifié.
Un écart significatif est :

$$R^2 - R_q^2 > \frac{q}{n-p-1}(1 - R^2)\mathcal{F}(1-\alpha; q, n-p-1)$$

Algorithmes de sélection

- pas à pas :
  - forward : à chaque étape on ajoute une variable, celle associée à la p-valeur de la statistique de Fisher est la plus petite (seuil 0.5)
  - backward : à chaque étape on retire une variable, celle associée à la p-valeur de la statistique de Fisher est la plus grande (seuil 0.1)
  - stepwise : mixte
- global : Furnival et Wilson
  Algorithme exhaustif qui cherche à optimiser un critère (R^2, R^2_a, C_p)

Inférence