Les méthodes de régression

Projet de recherche

Fried-Junior SABAYE
François BOUSSENGUI
Nicolas BARRÉ

October 6, 2020

Table des matières

Section 1

Moindres carrés ordinaires

Cadre général

- Considérons tout d'abord un modèle de régression linéaire simple :

$$y = \beta_0 + \beta_1 X + e$$

où :

- **y** est la variable à expliquer et **X** est la variable explicative du modèle. Les paramètres du modèle ou encore coefficients de régression, β_0 et β_1 sont respectivement l'ordonnée à l'origine et la pente de la droite de régression associée a cette équation. Enfin, **e** représente la différence entre les vraies valeur et les valeurs observées de la **y**. Pour des raisons d'inférences statistiques, on affirme que **e** est une variables aléatoire I.I.D tel que $\mathbf{e} \sim N(0, \sigma^2)$

- **y** est considérée comme une variable aléatoire avec :
  $\mathbb{E}(\mathrm{y}) = \beta_0 + \beta_1 \mathrm{X}$ et $\mathrm{var}(\mathrm{y}) = \sigma^2$
- Parfois **X** peut être considéré également comme une variable aléatoire. Dans ce cas, on prend en compte la moyenne et la variance conditionnelle de **y** sachant $\mathbf{X} = x$:
  $\mathbb{E}(\mathrm{y} \mid x) = \beta_0 + \beta_1 \mathrm{X}$ et $\mathrm{var}(\mathrm{y} \mid x) = \sigma^2$
- Les paramètre β_0, β_1 et σ^2 sont inconnus et **e** est inobservables.
- Afin de déterminer une estimation de ce paramètres, l'une des techniques couramment utilisée est celle des **moindres carrés ordinaires**.

Moindres carrés ordinaires

- On observe un échantillon de N ensembles d'observation (x_i, y_i) (i = 1, ..., N). Nous pouvons écrire l'équation précédente comme suit :

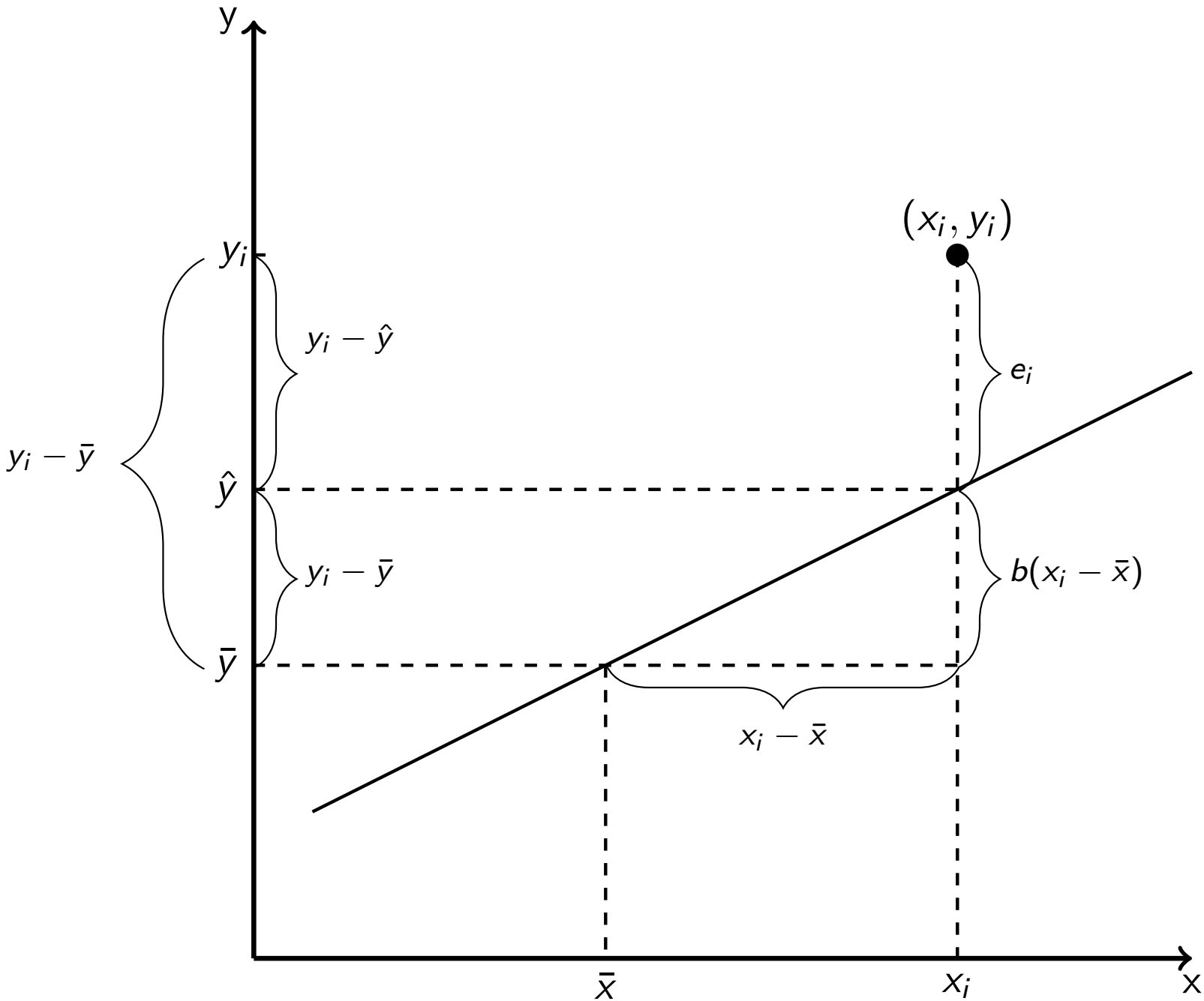
$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- Le principe de la méthode des moindres carrés ordinaires consiste à estimer les paramètres β_0 et β_1 de sorte que la somme des carrés de la différence verticale entre les observations et la droite de régression soit minimale.

- On minimise la quantité suivante afin d'avoir une estimation des coefficients :

$$S(\beta_0, \beta_1) = \sum_{i=1}^{N}(e_i^2) = \sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2$$

- En égalisant l'équation précédente à 0 on obient donc les solutions de β_0 et β_1 qui sont les estimateurs des moindres carrés ordinaires de β_0 et β_1. Ces estimateurs sont :
- $b_0 = \overline{y} - b_1\overline{x}$
- $b_1 = \frac{cov[X,Y]}{\sigma_x^2}$


y
(x_i, y_i)
y_i
$y_i - \hat{y}$
e_i
$y_i - \bar{y}$
$\hat{y}$
$y_i - \bar{y}$
$b(x_i - \bar{x})$
$\bar{y}$
$x_i - \bar{x}$
$\bar{x}$
x_i
x

Regression linéaire multiple

- Un modèle de régréssion linéaire multiple est une extension du modèle de régression linéaire simple dans le sens où il y a plusieurs variables explicatives.
- L'équation de régression pour la i-ème observation s'écrit :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

où :

- i = 1, ..., N correspond au nombre d'observations
- y_i est la i-ème observation de la variable dépendante y
- $x_{i,j}$ est la i-ème observation de la k-ème variable avec k = 0, ..., J
- ϵ_i : erreur du modèle (v.a.r) (part de la variabilité de Y qui n'est pas expliquée par le lien fonctionnel linéaire)

Régression linéaire multiple

- Pour tous les individus N, le modèle peut s'écrire sous la forme matricielle :

$$\mathbf{y} = \mathbf{X}\,\beta + \varepsilon$$

- **y** est un vecteur de dimensions [N x 1]
- X est une matrice de dimensions [N x (J + 1)]
- β est un vecteur de dimensions [(J+1) x 1]
- ε est un vecteur de dimensions [N x 1]

Régression linéaire multiple

Estimateurs

La méthode standard d'estimation des paramètres utilisée ici est la même que dans le cas de la regression linéaire simple à savoir : les **moindres carrés ordinaires** :

$$S(\beta_0,\ldots,\beta_p)=\sum_{i=1}^{n}(\epsilon_i{}^2)=\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_ij)^2=||\epsilon||^2$$

Après résolution, on à :

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Propriétés des estimateurs

- $\mathbb{E}[\boldsymbol{\beta}]=\boldsymbol{\beta}$ estimateur sans biais
- $\mathbb{V}[\boldsymbol{\beta}]=\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

Application sur Rstudio

La fonction lm()

- Package nécessaire : lm
- Définition du modèle :
  Modèle <- lm(Y$\sim$X$_1$, ..., X_n, dataframe, subset, weights,...)

Estimateurs de Theil - Sen

L'estimateur Theil-Sen, tel que défini par Theil (1950), d'un ensemble de points bidimensionnels (x_i, y_i) est la médiane des pentes $\frac{(y_j - y_i)}{(x_j - x_i)}$ déterminé par toutes les paires de points d'échantillonnage. Sen (1968) a étendu cette définition pour traiter le cas où deux points de données ont la même coordonnées x.

Cet estimateur se calcul efficacement et est insensible aux valeurs aberrantes. Il peut être nettement plus précis que les MCO dans le cadre la régression linéaire simple pour les données asymétriques et hétéroscédastiques, et rivalise bien avec les MCO pour les données normalement distribuées en termes de puissance statistique.

Theil - Sen

Considérons un modèle de régression linéaire simple :

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

Géométriquement, dans le but d'estimer la pente β_1, seuls deux points distincts $(x_i, y_i), (x_j, y_j)$ $(x_i \neq x_j)$ sont utiles.

Alternativement, avec deux points distincts, la somme des carrés des résidus est $(y_i - \beta_0 - \beta_1 x_i)^2 + (y_j - \beta_0 - \beta_1 x_j)^2$. Cette quantité est minimisée quand β_0, β_1 satisfont les équations :

$$y_i - \beta_0 - \beta_1 x_i = 0, \qquad y_j - \beta_0 - \beta_1 x_j = 0$$

Les solutions sont :

- $\hat{\beta}_{0\,i,j} = y_i - \hat{\beta}_{1\,i,j} x_i$
- $\hat{\beta}_{1\,i,j} = \frac{y_i - y_j}{x_i - x_j}$

Ce sont les estimateurs des moindres carrés. Un estimateur robuste de la

Theil - Sen

Un estimateur robuste de la pente est la médiane de ces estimations par moindres carrés :

$$\widetilde{\beta_1} = \text{Med}\ \{\ \hat{\beta}_{1i,j} = \frac{y_i - y_j}{x_i - x_j}\ :\ x_i \neq x_j,\ 1 \leq \text{i} \leq \text{j} \leq \text{n}\ \}$$

De même, β_0 (ordonnée à l'origine) peut être estimée, sous certaines conditions, par la médiane des estimations par moindres carrés :

$$\widetilde{\beta_0} = \text{Med}\ \{\ \hat{\beta}_{0i,j} = \frac{y_j x_i - y_i x_j}{x_i - x_j}\ :\ x_i \neq x_j,\ 1 \leq \text{i} \leq \text{j} \leq \text{n}\ \}$$

Application sur Rstudio

La fonction mblm()

- Package nécessaire : mblm
- Définition du modèle :
  Modèle <- mblm(Y$\sim X_1, ..., X_n$, dataframe, repeated = TRUE)

Section 2

Régression Orthogonale

Cadre général

Le principe des moindres carrés dans la régression orthogonale est la minimisation de la distance perpendiculaire au carré entre les points et la droite de régression afin d'obtenir une estimation des coefficients de régression.

Nous avons le modèle suivant :

$$Y_i = \beta_0 + \beta_1 X_i$$

qui définit

$$E_i = Y_i - \beta_0 - \beta_1 X_i = 0 \quad (1)$$

Régression orthogonale

Les coefficients de régression sont obtenus en minimisant la distance perpendiculaire au carré entre les points $(x_i, y_i)(i = 1, ..., N)$:

$$d_i^2 = (X_i - x_i)^2 + (Y_i - y_i)^2 \quad (2)$$

L'objectif étant de minimiser $\sum_{i=1}^{N} d_i^2$ pour obtenir les estimations de β_0 et de β_1.

Les coefficients de régression sont obtenus en minimisant (2) sous la contrainte (1) en utilisant la méthode du multiplicateur de Lagrange.

La fonction de Lagrange étant :

$$L = \sum_{i=1}^{N} d_i^2 \text{ - } 2 \sum_{i=1} \lambda_i \, E_i$$

où $\lambda_1, ..., \lambda_i$ sont les multiplicateurs de Lagrange.

Après résolution nous obtenons une estimation de régression orthogonale de β_0 et $\beta_1$tel que :

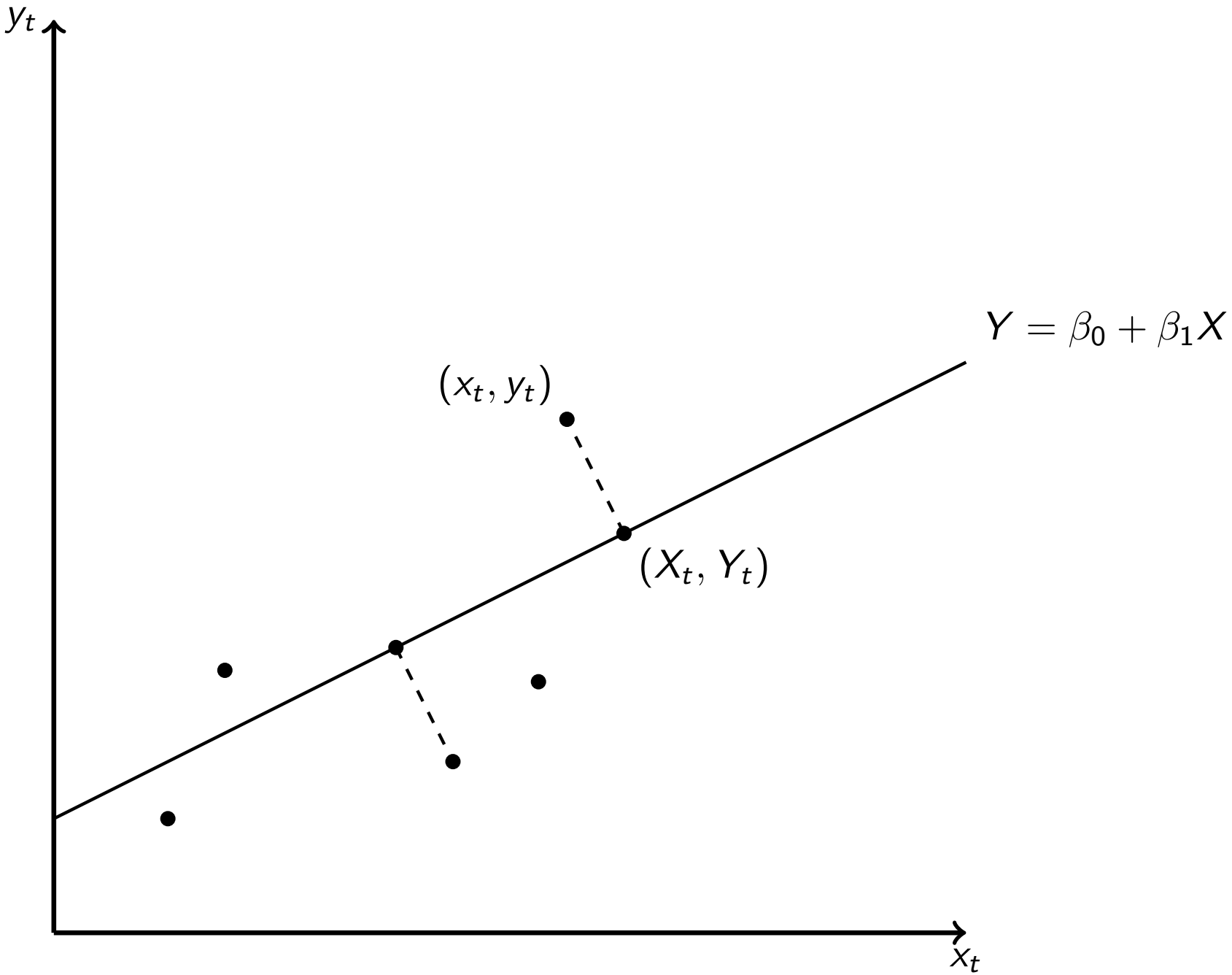
$$\hat{\beta}_{0\,OR} = \bar{y} - \hat{\beta}_{1\,OR}\bar{x}$$

et

$$\hat{\beta}_{1\,OR} = \frac{(SYY - SXX) + sgn(SXY)\sqrt{(SXX - SYY)^2 + 4SXY}}{2SXY}$$

Où :

- sgn(SXY) est le signe de (SXY)
- SXX et SYY respectivement Var(X) et Var(Y)
- SXY = cov(XY)


y_t
$Y = \beta_0 + \beta_1 X$
(x_t, y_t)
(X_t, Y_t)
x_t

Régression de DEMING

Les modèles d'erreur de mesure supposent que la variable réponse et qu'une ou plusieurs des variables prédictives font l'objet d'erreur de mesure.

De manière générale, nous avons :

$$Y = y_{true} + \varepsilon \quad (1)$$
$$W = X + U \quad (2)$$

- Y = Variable à expliquer observée, $Y = y_{true} + \varepsilon$, var$(\varepsilon) = \sigma_\varepsilon^2$
- y_{true} = "Vraie" valeur de la variable à expliquer
- W = Prédicteur observé, $W = X + U$, var(U) $= \sigma_u^2$
- X = "Vraie" valeur du prédicteur
- ε et U sont indépendants

Régression de DEMING

- En combinant (1) et (2) nous avons le modèle suivant :

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

- Un modèle de régression de DEMING requiert la connaissance du ratio des variances :

$$\eta = \frac{var(Y|X)}{var(W|X)} = \frac{\sigma_\varepsilon^2}{\sigma_u^2}$$

Régression de DEMING

- L'estimateur de la régression de DEMING s'obtient en minimisant la quantité :

$$\sum_{i=1}^{N} \{(Y_i - \beta_0 - \beta_1 X_i)^2/\eta + (W_i - X_i)^2\} \quad (3)$$

- β_0, β_1 , X_1, ... X_n sont des inconnus.
- Si $\eta = 1$, alors (3) sera égale à la distance orthogonale de $(Y_i, W_i)_{i=1}^n$ de la droite $(\beta_0 + \beta_1 X_i, X_i)_{i=1}^n$.
- Si $\eta \neq 1$ alors (3) est une distance orthogonale pondérée.

- L'estimateur de la régression de DEMING est donc :

$$\beta_1(\text{OR}) = \frac{s_y^2 - \eta s_w^2 + \{(s_y^2 - \eta s_w^2)^2 + 4\eta s_{wy}^2\}^{1/2}}{2s_{wy}}$$

Avec :

- var(Y) $= s_y^2$
- var(W) $= s_w^2$
- Cov(WY) $= s_{wy}$

Régression Quantile

Un des problèmes avec la régression linéaire classique est la sensibilité des coefficients aux valeurs extrèmes. Ce fait est du à la dépendance des estimateurs à la **moyenne empirique**.

$$\alpha = \overline{Y} - \beta\overline{X}$$
$$\beta = \frac{\Sigma_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma_{i=1}^{N}(X_i - \overline{X})}$$

Une manière alternative d'estimer un modèle linéaire est d'utiliser des estimateurs β qui ne dépendent pas de la moyenne empirique comme les estimateurs MCO. La regression quantile permet de régresser sur des quantiles précis des variables explicatives.

Pour un modèle $Y = X\beta + \epsilon$,

La fonction objectif à minimiser pour une regression au θ^e quantile est :

$$Min\{\sum_{Y_i \leq \beta X_i} \theta |Y_i - \beta X_i| + \sum_{Y_i > \beta X_i} (1-\theta)|Y_i - \beta X_i|\}$$

- On ne minimise pas le carré des résidus mais la valeur absolue de l'erreur, pondérée par la pénalité θ
- On peut décider de régresser un échantillon sur plusieurs quantiles pour mieux rendre compte de la dispersion sur chaque partie de la population.
- Dans des cas où la variance du terme d'erreur n'est pas constante (hétéroscedasticité), cette méthode fournit de meilleures estimations que les MCO.
- On peut résoudre ce problème en utilisant la programmation linéaire.

Application sur Rstudio

La fonction rq()

- Package nécessaire : quantreg
- Définition du modèle :
  Modèle <- rq(Y$\sim X_1, ..., X_n$, dataframe, tau = c(0.1,0.5,0.9))

Dans ggplot

- L'argument geom$_q$*uantile*() :

ggplot() + *geom*$_q$*uantile*(*quantiles* = *c*(...))

Section 3

Régression non paramétrique

Cadre général

Dans un modèle de régression non-paramétrique, la fonction de lien entre **X** et **Y** n'a pas de forme explicite et ne peut pas s'écrire en fonction d'un nombre réduit de paramètres. On cherche :

$$y = f(x) + \epsilon \text{ , où } E(Y|X = x) = f(x)$$

Avec une approche non paramétrique on aboutit à :

- une relation graphique entre **X** et **Y**
- des estimateurs (smoothers) beaucoup plus souples grâce au peu d'hypothèses que nous gardons
- Il n'existe pas de forme analytique de la fonction de lien f (x)

Régression kernel

- La régression avec lissage par opérateur à noyau ou régression kernel cherche à estimer la fonction $f(x_i)$ en tout point x_1, x_2 ,., x_N. Pour cela on utilise communément le lissage par opérateur à noyau ou kernel smoother (Nadaraya, 1964 et Watson, 1964).

- L'estimateur à noyau (kernel estimate) de la fonction de lien évaluéé au point x_0, noté $f(x_0)$, est défini par :

$$f(x_0) = \sum_{i=1}^{N} w_i(x_0) y_i$$

avec :

$$w_i(x_0) = \frac{K(\frac{x_i - x_0}{\lambda})}{\sum_{i=1}^{N} K(\frac{x_i - x_0}{\lambda})}$$

Où K(.) désig,e une fonction kernel, $\lambda > 0$ un paramètre de lissage (bandwidth parameter) et N la taille de l'échantillon utilisée pour l'estimation

- **Remarque 1** : La fonction de lien évaluée au point de x_0 est donc définié comme une somme pondérée des observations y_i dont les poids $w_i(x_0)$ dépendent de x_0.
- **Remarque 2** : La fonction $w_i(x_0)$ définit le poids qui doit être attribué au couple d'observations (x_i, y_i) dans la valeur de la fonction de lien évaluée au point d'abscisse x_0. Généralement, plus les points x_i sont proches de x_0, plus le poids sera important : $w_i(x_0)$ est donc décroissante dans la disance $|x_0 - x_i|$

Ces poids dépendent de fonction kernel (ou opérateur à noyau) qui correspond tout simplement à des fonctions de densité de probabilté.

Une fonction de kernel $K(\frac{x_i - x_0}{\lambda}) = \mathrm{K(u)}$ vérifient les propriétés suivantes:

- (i) $\mathrm{K(u)} \geq 0$
- (ii) K(u) est normalisé de sorte que : $\int \mathrm{K(u)} du = 1$
- (iii) K(u) atteind son maximum en 0 lorsque $x_i = x_0$ et décroît avec la distance $|x_0 - x_i|$.
- (iv) K(u) est symétrique : le kernel ne dépende que de la distance $|x_0 - x_i|$ et non du signe de $x_0 - x_i$.

Régressions locales

Un des problèmes essentiels avec la régression Kernel réside dans le manque de robustesse de ces estimateurs pour les valeurs extrêmes de $\mathbf{X}$. Une solution alternative, plus robuste pour les valeurs extrêmes consiste en l'utilisation de régression locales.

Le principe général d'une **régression locale** est de postuler que la fonction de lien $f(x_0)$ évaluée au point x_0 peut être approximé par la valeur d'une fonction paramétrique évalué localement au voisinage $N(x_0)$ du point de référence x_0.

La **procédure LOESS** attribue des poids selon une fonction de type tri cubique en fonction de la distance au centre de classe.

La fonction de pondération généralement utilisée pour effectuer une régression locale est une fonction cubique pondérée :

$$w(x) = (1 - |x|^3)^3 I[|x| < 1]$$