Semaine 4 : Plusieurs façons de régresser !

1 Vous avez dit régression ?

1.1 Objectifs et contexte

Objectifs

- Vérifier si une relation linéaire existe entre les variables aléatoires continues Y (variable réponse) et X (variable explicative).
- Mesurer la force du lien linéaire qui unit les variables.
- Trouver la meilleure droite exprimant la relation entre les 2 variables à partir de n paires d'observations indépendantes $(X_i,\ Y_i)$.
- Prédire la valeur de Y pour un X donné, et évaluer la précision de ces prédictions.

Contexte
On veut mettre en relation deux variables aléatoires continues, X et Y.

On basera notre analyse sur n couples de données : $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$,
dont les valeurs observées seront notées $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.

Pour obtenir ces couples de données, deux scénarios d'échantillonnage peuvent être envisagés.

Scénario d'échantillonnage 1 :
1. Les valeurs de X sont fixées d'avance

X est une variable contrôlable, et seules les valeurs de Y sont réellement mesurées.

Exemple :

Un chimiste veut mesurer l'effet de la température sur la vitesse de réaction lors du mélange de deux liquides. Il peut réaliser son expérience à plusieurs reprises et contrôler la température des liquides, en la fixant à différentes valeurs. Il mesure chaque fois la vitesse de réaction.

Scénario d'échantillonnage 2 :
2. Les valeurs de X et de Y sont aléatoires

X et Y sont des variables dont les mesures sont prises simultanément sur des individus (unités expérimentales) sélectionnés aléatoirement.

Exemple :

Une chercheuse en obstétrique veut étudier le lien entre le poids de la mère et le poids du nouveau-né. Elle sélectionne des femmes enceintes au hasard, et mesure le poids du bébé et de la mère au moment de l'accouchement.

1.2 Le modèle linéaire

Le modèle de régression linéaire simple
On suppose que X et Y sont reliées par la droite $Y = \beta_0 + \beta_1 X$, et que les observations dévient un peu de ce modèle par une erreur aléatoire ε, ce qui s'écrit (pour un point en particulier) :

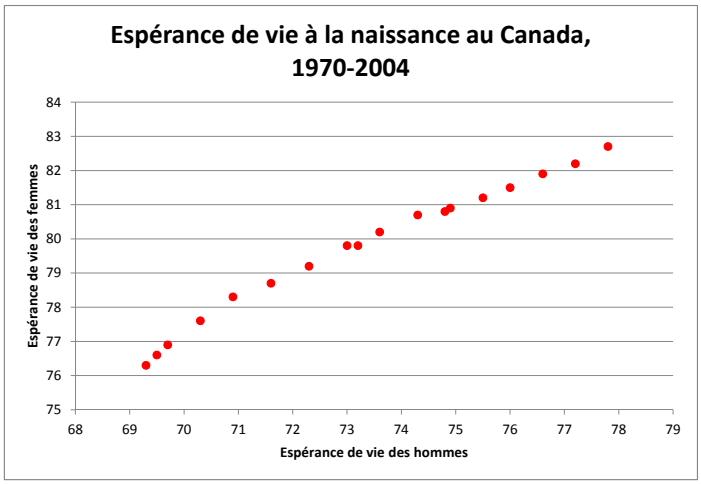
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{où} \quad i = 1, \ldots, n,$$

$$\begin{aligned} Y_i &= i^e \text{ mesure de la variable } Y, \\ \beta_0 &= \text{ordonnée à l'origine}, \\ \beta_1 &= \text{pente de la droite de régression}, \\ X_i &= i^e \text{ mesure de la variable } X, \\ \varepsilon_i &= \text{perturbation due au hasard ou à des variables autres que X} \\ &\quad \text{(erreur du modèle)}. \end{aligned}$$

Un exemple
On veut mettre en relation l'espérance de vie à la naissance des hommes et celle des femmes du Canada. On dispose de données compilées de 1970 à 2004.

Année	Hommes	Femmes
1970	69,3	76,3
1972	69,5	76,6
1974	69,7	76,9
1976	70,3	77,6
1978	70,9	78,3
1980	71,6	78,7
1982	72,3	79,2
1984	73,0	79,8
1986	73,2	79,8
1988	73,6	80,2
1990	74,3	80,7
1992	74,8	80,8
1994	74,9	80,9
1996	75,5	81,2
1998	76,0	81,5
2000	76,6	81,9
2002	77,2	82,2
2004	77,8	82,7

Un exemple



2 Estimation des paramètres β_0 et β_1

Estimation des paramètres β_0 et β_1
Puisque les points $(x_1, y_1), ...(x_n, y_n)$ ne sont en général pas alignés, il faut trouver la droite qui représente le mieux la relation entre X et Y.
Nous considérerons trois approches :

1. Méthode de Mayer : trouvons les deux points les plus ”représentatifs” et lions-les.

2. Méthode médiane-médiane : on utilise les points médians de l'échantillon séparé en trois.

3. Méthode des moindres carrés : on veut minimiser la distance totale entre les points et la droite.

2.1 Méthode de Mayer

1. Méthode de Mayer
Antoine Falguerolles (2009, Université de Toulouse) présente la méthode de Mayer comme suit :

Cherchant à résoudre un système d'équations linéaires numériquement incompatibles, l'astronome Tobias Mayer (...) propose de sommer (ou moyenner) ces équations par groupe en définissant autant de groupes disjoints d'observations qu'il y a de coefficients à estimer. (...) La méthode, publiée par Mayer (...) exige donc qu'une partition soit fournie a priori mais son auteur ne propose pas de procédure générale permettant de guider le choix de cet élément décisif.

1. Méthode de Mayer

1.

2.

1. Méthode de Mayer (suite)

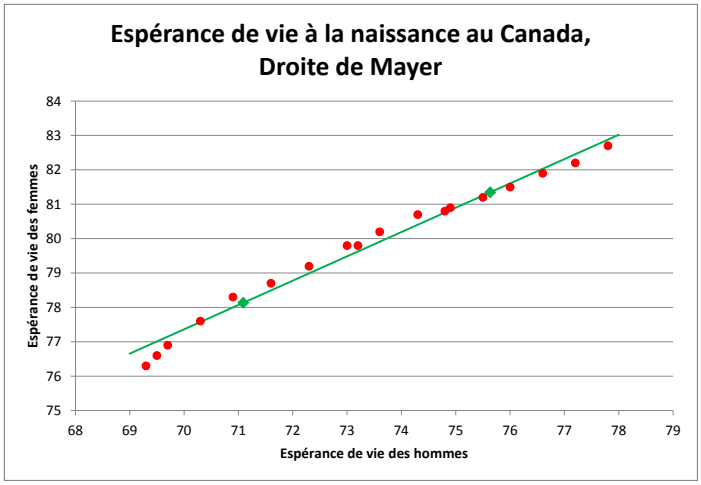
3.

Tobias Mayer : 1723-1762
Astronome allemand en charge de l'Observatoire de Göttingen. Calcula avec précision les mouvements de la Lune. Professeur de mathématiques à l'Université de Göttingen. (Tout comme Gauss (1807), Dirichlet(1831), Riemann(1859))

Exemple
Déterminer l'équation de la droite de Mayer.

Année	Hommes	Femmes
1970	69,3	76,3
1972	69,5	76,6
1974	69,7	76,9
1976	70,3	77,6
1978	70,9	78,3
1980	71,6	78,7
1982	72,3	79,2
1984	73,0	79,8
1986	73,2	79,8
1988	73,6	80,2
1990	74,3	80,7
1992	74,8	80,8
1994	74,9	80,9
1996	75,5	81,2
1998	76,0	81,5
2000	76,6	81,9
2002	77,2	82,2
2004	77,8	82,7

Exemple



Caractéristiques de la méthode de Mayer

1.

2.

3.

2.2 Méthode médiane-médiane (Med-med)

2. Méthode médiane-médiane (Med-med)

1. L'échantillon est partitionné en trois sous-ensembles à peu près égaux, suivant les valeurs de x, de telle sorte que le premier et le dernier groupes contiennent le même nombre de points.

2. On calcule le *point médian* de chacun des trois sous-ensembles.

3. On relie les points médians des deux parties extrêmes.

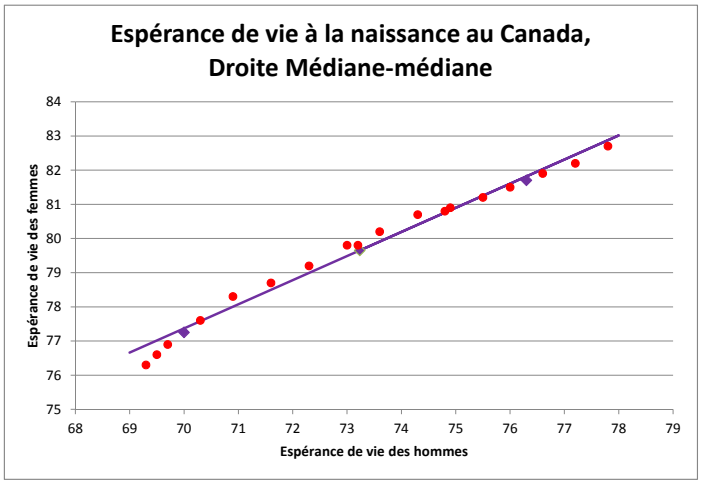
4. On déplace cette droite de façon parallèle jusqu'à ce qu'elle passe par le *point moyen des trois couples de médianes*.

Exemple
Déterminer l'équation de la droite médiane-médiane.

Année	Hommes	Femmes
1970	69,3	76,3
1972	69,5	76,6
1974	69,7	76,9
1976	70,3	77,6
1978	70,9	78,3
1980	71,6	78,7
1982	72,3	79,2
1984	73,0	79,8
1986	73,2	79,8
1988	73,6	80,2
1990	74,3	80,7
1992	74,8	80,8
1994	74,9	80,9
1996	75,5	81,2
1998	76,0	81,5
2000	76,6	81,9
2002	77,2	82,2
2004	77,8	82,7

Exemple

Exemple



Caractéristiques de la méthode Med-med

1.

2.

3.

2.3 Méthode des moindres carrés

3. Méthode des moindres carrés
Pour estimer β_0 et β_1, on minimise la somme des carrés des erreurs, i.e. des écarts verticaux entre les observations Y_i et le point correspondant sur la droite de régression $\beta_0 + \beta_1 X_i$:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - [\beta_0 + \beta_1 X_i])^2$$

Il suffit donc de dériver cette équation successivement par rapport à β_0 et β_1, d'égaler les dérivées partielles à 0 et de résoudre le système obtenu pour $\hat{\beta}_0$ et $\hat{\beta}_1$.

Estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Notations

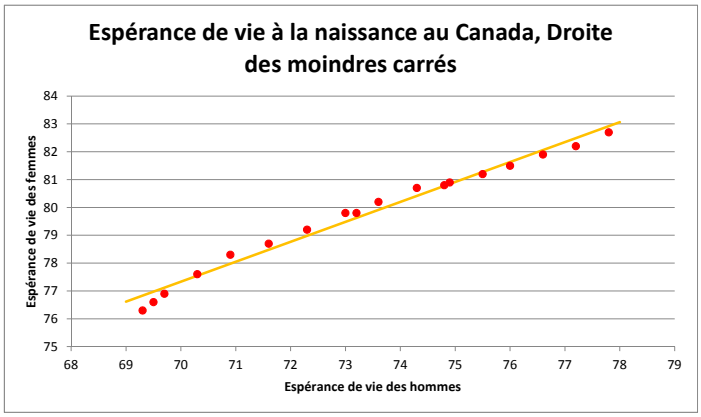
$$S_{xx} = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = (n-1)s_X^2$$

$$S_{yy} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 = (n-1)s_Y^2$$

$$S_{xy} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}$$

Exemple

Exemple



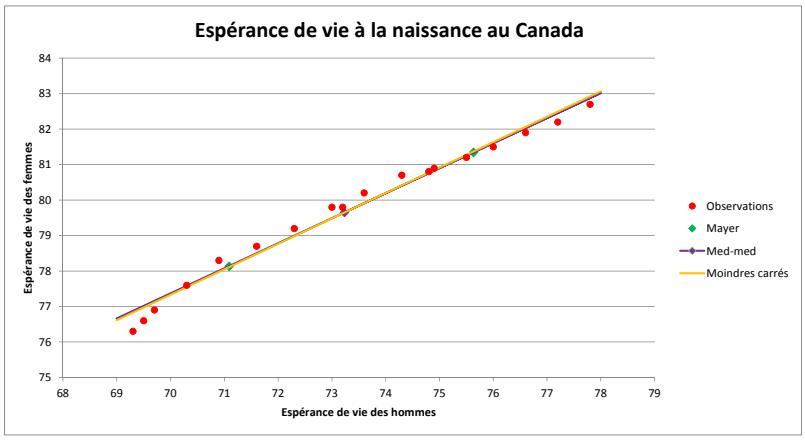
Caractéristiques de la méthode des moindres carrés

1.

2.

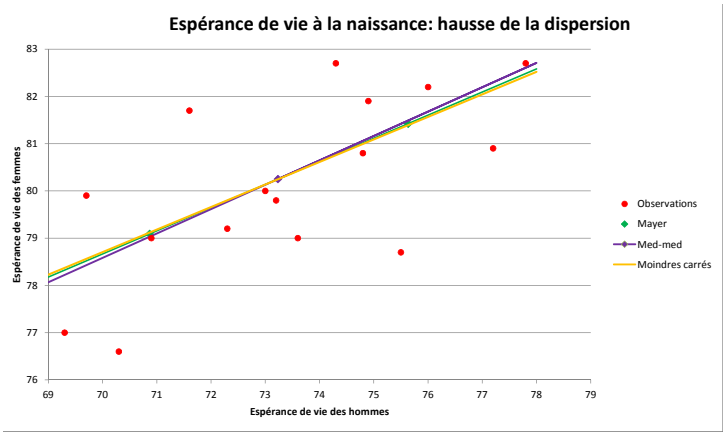
3.

Comparaison des trois droites

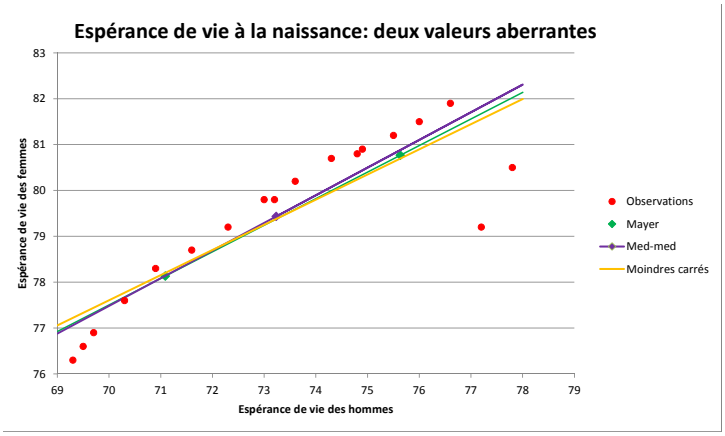


Données trafiquées : hausse de la dispersion

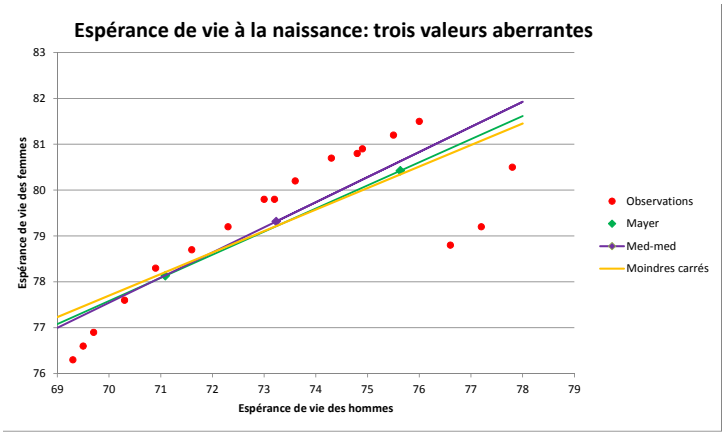
On voit que les droites diffèrent peu lorsque le nuage de points est relativement linéaire sans valeurs aberrantes.



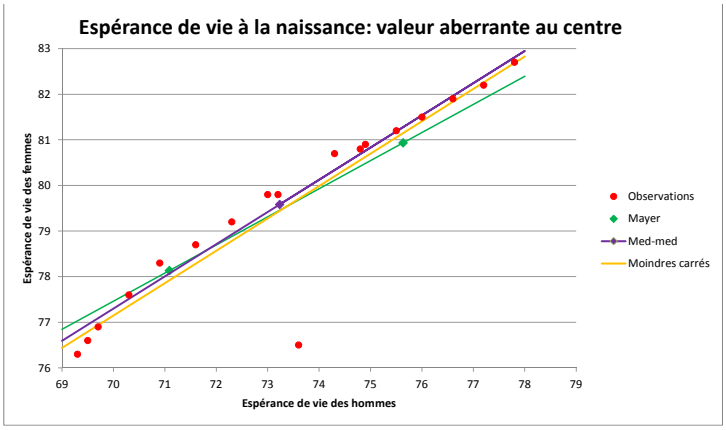
Données trafiquées : deux valeurs aberrantes
Les trois droites sont à côté de la partie linéaire du nuage de points. La med-med est la moins affectée.



Données trafiquées : trois valeurs aberrantes
Toutes les droites sont à côté de la partie linéaire du nuage de points. C'est quand même la med-med qui est la moins affectée.



Données trafiquées : valeur aberrante au centre
Moindres carrés : On voit que la pente n'est pas affectée, mais que l'ordonnée à l'origine l'est. Mayer : La plus affectée, à cause du calcul du 2^e point moyen. Med-med : n'est affectée que si la valeur aberrante est impliquée dans l'un des points médians.


Espérance de vie à la naissance: valeur aberrante au centre
Espérance de vie des femmes
Espérance de vie des hommes
83
82
81
80
79
78
77
76
69
70
71
72
73
74
75
76
77
78
79
Observations
Mayer
Med-med
Moindres carrés