La regression linéaire

Dr F. Moreau

03/07/2024

1 Présentation

La régression linéaire est, à la base, une méthode statistique qui analyse la pertinence d'une éventuelle relation de dépendance linéaire entre une variable dépendante, y, et une ou plusieurs variables indépendantes, les x_i.

Dans le contexte du machine learning, les variables dites indépendantes sont appelées les caractéristiques et la varible indépendante est la cible. Par exemple, y pourrait être la valeur pécuniaire d'une maison et les caractéristiques x_i seraient, le nombre de pièces de la maison, la superficie de son terrain, la proximité de certains services, son année de construction, des informations liées à son emplacement, etc. On cherche à établir et à évaluer un modèle linéaire qui lie les caractéristiques à la cible. Ce modèle s'exprime

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i \, x_i + \epsilon \tag{1}$$

où les coefficients β_i sont les $n+1$ paramètres du modèle. ϵ est le terme erreur, le terme résiduel, représentant la part de la cible y qui ne peut pas être expliquée par le modèle linéaire. Si le modèle linéaire est totalement pertinent, la quantité ϵ ne représente qu'une variabilité individuelle au sens statistique et doit posséder une distribution aléatoire de moyenne nulle.

Etablir un modèle linéaire consiste à déterminer les valeurs optimales des paramètres β_i, c'est-à-dire ceux qui minimisent le terme résiduel. Reprenons l'exemple des maisons. Imaginons que, pour une maison particulière, la numéro k, les valeurs des caractéristiques soient notées x_{ki}. La valeur prédite par notre modèle linéaire est notée $\hat{y}_k$. Nous avons

$$\hat{y}_k = \beta_0 + \sum_{i=1}^{n} x_{ki} \, \beta_i \tag{2}$$

Cette valeur sera très certainement différente de la valeur affichée de la maison qui nous notons y_k. La différence est le terme ϵ. Nous voulons minimiser cet écart résiduel. En

pratique, nous disposons de données sur un grand nombre m de maisons (des milliers voire beaucoup plus). Pour chacune d'entre elles, il y a un écart entre la valeur prédite par le modèle, $\hat{y}_k$, et la valeur effectivement observée y_k et ce pour k allant de 1 jusque m. Ce que nous cherchons à faire est de minimiser globalement l'ensemble de ces écarts résiduels. La façon la plus courante (de loin) d'y parvenir est de minimiser la somme des carrés de ces écarts résiduels (SCER). Nous cherchons donc à minimiser la quantité

$$SCER = \sum_{k=1}^{m} (y_k - \hat{y}_k)^2 = \sum_{k=1}^{m} \left(y_k - \left(\beta_0 + \sum_{i=1}^{n} x_{ki}\, \beta_i \right) \right)^2 \tag{3}$$

où m est le nombre d'items ou d'individus (au sens statistique), c'est-à-dire le nombre de maisons dans notre exemple. La relation 2 nous donne une interprétation élémentaire des paramètres. Le paramètre β_0 est l'estimation de la valeur cible lorsque toutes les caractéristiques sont nulles. Les paramètres β_i avec $i = 1, \cdots, n$ sont les pentes, c'est-à-dire la variation de la valeur cible lorsque la caractéristique i se modifie d'une unité, les autres caractéristiques étant constantes.

La validité de l'analyse mathématique du problème de régression linéaire repose sur certaines hypothèses :

- Linéarité : la valeur cible y dépend linéairement des variables indépendantes, les x_i
- Indépendance : les résidus ϵ pour chaque individu doivent être indépendants
- Homoscédasticité : les résidus doivent être de même variance
- Normalité : les résidus suivent une loi de distribution normale
- Collinéarité : les variables indépendantes ne doivent pas être linéairement dépendantes les unes des autres (évidemment) et les cross sorrélations ne doivent pas être trop élévées. Ce point sera abordé dans une section ultérieure.

Les 3 conditions concernant les résidus sont vérifiés si les résidus sont les réalisations indépendantes d'une même variable aléatoire normale.

2 Les équations normales

Le problème de minimisation de la SCER (relation 3) possède une solution analytique que nous allons aborder. Pour cela, nous introduisons une formulation vectorielle du problème. Toutes les valeurs des n caractéristiques pour chacun des m individus (items) vont être rangées dans un matrice X de m lignes et n colonnes. Chaque ligne de cette matrice contient donc les caractéristiques d'un individu. On adjoint sur la gauche de cette matrice une colonne de 1. Au final X a donc m lignes et $n+1$ colonnes. Les $n+1$ paramètres forment un vecteur colonne $\vec{\beta}$ à $n+1$ composantes. La première de ces composantes est β_0.

La relation 2 peut donc s'écrire pour tous les individus simultanément

$$\vec{\hat{y}} = X\vec{\beta} \tag{4}$$

Les m composantes du vecteur $\vec{\hat{y}}$ sont les valeurs prédites par le modèle pour chacun des m individus. De la même façon, les valeurs cibles effectivement observées peuvent être rangées dans un vecteur $\vec{y}$ à m composantes. La relation 1 peut alors s'écrire

$$\vec{y} = X\vec{\beta} + \vec{\epsilon}$$

où le vecteur $\vec{\epsilon}$ contient les termes résiduels de chaque individu. La SCER que nous cherchons à minimiser est, d'après 3, simplement la norme au carré de la différence entre les vecteurs $\vec{y}$ et $\vec{\hat{y}}$

$$\begin{aligned} \text{SCER} &= \left\|\vec{y} - \vec{\hat{y}}\right\|^2 \\ &= \left\|\vec{y} - X\vec{\beta}\right\|^2 \\ &= \left(\vec{y} - X\vec{\beta}\right)^T \cdot \left(\vec{y} - X\vec{\beta}\right) \\ &= \|\vec{y}\|^2 - \left(X\vec{\beta}\right)^T \cdot \vec{y} - \vec{y}^T \cdot \left(X\vec{\beta}\right) + \left(X\vec{\beta}\right)^T \cdot \left(X\vec{\beta}\right) \end{aligned} \tag{5}$$

Grâce aux relations 33, 34 et 35 de l'annexe 6.1, nous pouvons réécrire cette dernière ligne

$$\text{SCER} = \|\vec{y}\|^2 - 2\vec{\beta}^T \cdot \left(X^T\vec{y}\right) + \vec{\beta}^T \cdot \left(X^T X\vec{\beta}\right) \tag{6}$$

Nous cherchons la position dans l'espace des paramètres $\mathbb{R}^{n+1}$ qui minimise la SCER. Vu que la SCER est une fonction continument dérivable par rapports aux paramètres, au minimum de cette fonction, ses dérivée partielles par rapport aux paramètres sont nulles.

Au départ de la relation 6, et en utilisant les résultats 36 et 37 de l'annexe 6.1 nous pouvons calculer le gradient de la SCER dans l'espace des paramètres

$$\vec{\nabla}_{\vec{\beta}} SCER = -2X^T\vec{y} + 2X^T X\vec{\beta} \tag{7}$$

L'annulation des dérivées premières au minimum mène donc à

$$X^T X\vec{\beta} = X^T\vec{y} \tag{8}$$

La relation 8 est un système d'équations linéaires dont les solutions sont les paramètres β_i. Matrice du système, $X^T X$, est de taille $(n+1)\times(n+1)$. Il ne faut pas espérer résoudre ce système par une méthode classique car $X^T X$ est souvent très mal conditionnée. La résolution doit se faire par la technique SVD (Singular Value Decomposition). L'obtention

de la matrice du système nécessite $(n+1)^2 m$ multiplications et la résolution par SVD nécessite environ $12(n+1)^3$ multiplications. Les valeurs de n et m ne peuvent donc pas dépasser quelques milliers.

Si nous utilisons l'égalité 8 dans 6, cela donne

$$\begin{aligned}
\text{SCER} &= \|\vec{y}\|^2 - \vec{\beta}^T \cdot \left(X^T \vec{y}\right) \\
&= \|\vec{y}\|^2 - \left(X\vec{\beta}\right)^T \cdot (\vec{y}) \text{ en utilisant 33} \\
&= \|\vec{y}\|^2 - \left(X\vec{\beta}\right)^T \cdot \left(X\vec{\beta}\right) \\
&= \|\vec{y}\|^2 - \vec{\hat{y}}^T \cdot \vec{\hat{y}} \\
&= \|\vec{y}\|^2 - \left\|\vec{\hat{y}}\right\|^2
\end{aligned} \tag{9}$$

La minimisation des écarts résiduels revient donc à rendre la norme de $\vec{\hat{y}}$ aussi proche que possible de la norme de $\vec{y}$. Vu que $\vec{\hat{y}} = X\vec{\beta}$, la relation 8 nous dit aussi que

$$X^T \vec{\hat{y}} = X^T \vec{y} \tag{10}$$

Il faut voir X^T comme un opérateur de projection dans un sous-espace engendré par les vecteurs propres de X. Les projections dans ce sous-espace de $\vec{\hat{y}}$ et de $\vec{y}$ sont identiques.

Les relations 9 et 10 racontent en fait la même chose. Nous voulons rendre $\vec{\hat{y}}$ aussi proche que possible de $\vec{y}$ mais $\vec{\hat{y}}$ est coincé dans le sous-espace vectoriel engendré par les vecteurs propres de X. Le mieux que nous puissions faire est de prendre $\vec{\hat{y}}$ comme étant la projection de $\vec{y}$ sur ce sous-espace. Ce faisant, nous rendons minimale la distance entre ces deux vecteurs. C'est ce qui est exprimé par 9.

3 La fonction coût

Une alternative souvent employée est la minimisation par une technique itérative comme la méthode du gradient descent. Le principe de cette technique est abordé dans un autre document mais elle utilise le gradient de la fonction à minimiser. Or ce gradient a été calculé en 7. Pour faciliter la comparaison avec d'autres approches de machine learning on définit comme fonction à minimiser une fonction de coût notée J avec

$$J = \frac{1}{2m} SCER$$

Dès lors, le gradient de la fonction coût est

$$\vec{\nabla}_{\vec{\beta}} J = \frac{1}{m} X^T \left(X\vec{\beta} - \vec{y}\right) \tag{11}$$

4 Le coefficient de détermination

Pour juger de l'adéquation entre des données et un modèle de régression, on utilise le coefficient de détermination noté R^2. Nous allons voir la signification de ce coefficient et la façon dont on le calcule. Tout d'abord, intéressons-nous à la variance des valeurs cibles, les y. Nous allons voir que cette variance peut se séparer en deux parties très spécifiques.

$$\mathrm{Var}(y) = \frac{1}{m}\sum_{k=1}^{m}(y_k - \bar{y})^2$$

C'est la somme des carrés des écarts totaux (SCET) divisée par le nombre de points. Dans cette somme, on peut remplacer les y_k par $y_k - \hat{y}_k + \hat{y}_k$. Nous écrivons alors

$$\begin{aligned} SCET &= \sum_{k=1}^{m}\left((y_k - \hat{y}_k) + (\hat{y}_k - \bar{y})\right)^2 \\ &= \sum_{k=1}^{m}(y_k - \hat{y}_k)^2 + \sum_{k=1}^{m}(\hat{y}_k - \bar{y})^2 + 2\sum_{k=1}^{m}(y_k - \hat{y}_k)(\hat{y}_k - \bar{y}) \qquad (12) \\ & \qquad (13) \end{aligned}$$

Le premier terme du membre de droite est la somme des carrés des écarts entre les valeurs cibles et les valeurs théoriques (les ordonnées fournies par le modèle linéaire). Il s'agit donc Ce terme se nomme la somme des carrés des écarts résiduels (SCER). Le second terme est la somme des carrés des écarts entre les valeurs théoriques et la moyenne des y_k. Il se nomme la somme des carrés des écarts factoriels (SCEF). Le troisième terme est nul comme il est montré à l'annexe 6.2. On a donc

$$SCET = SCER + SCEF$$

La SCET représente la totalité de la dispersion, de la variabilité, des ordonnées y_k. De façon générale, les différences entre les valeurs cibles sont essentiellement dues aux caractéristiques des items. Reprenons l'exemple des maisons, SCET représente la variabilité totale des prix des maisons. Or, les différences de prix sont essentiellement dues aux caractéristiques telles que le nombre de pièces de la maison, la superficie de son terrain, la proximité de certains services, son année de construction, des informations liées à son emplacement, etc. Ce facteurs expliquent, à eux seuls, une grande partie de la variabilité des y_k. La partie de la variabilité expliquée par les caractéristiques des items est la SCEF. La SCER est donc la partie de la variabilité qui n'est pas explicable par les caractéristiques. Elle est soit due aux fluctuations du « hasard », soit due à d'autres facteurs noms pris en considération.

Le modèle est d'autant meilleur qu'il explique une plus grande part de la variabilité des y_k. Chercher un bon modèle revient donc à rendre la SCEF presqu'aussi grande que

SCET. On définit ainsi le coefficient de détermination R^2 par

$$R^2 = \frac{SCEF}{SCET} \tag{14}$$
$$= \frac{SCET - SCER}{SCET}$$
$$= 1 - \frac{SCER}{SCET} \tag{15}$$

R^2 se situe donc toujours entre 0 et 1 : $0 \leq R^2 \leq 1$. Plus R^2 est proche de 1, meilleur est le modèle. Donc, pour trouver un bon modèle, on cherche à réduire la SCER. C'est exactement ce que l'on a fait plus haut.

Vu que $SCET = m\text{Var}(y)$, on a ,à partir de 15,

$$SCER = m\text{Var}(y)(1 - R^2) \tag{16}$$

5 Regression linéaire sur un nuage de points à 2D

Dans ce cas, les items sont des points dans un espace à deux dimensions. Ils n'ont qu'une seule caractéristique : leur abscisse. Il n'est donc plus nécessaire de différencier les caractéristiques par un indice. On essaye de concevoir un modèle qui prédise au mieux l'ordonnée des points. Les ordonnées prédites par le modèle sont les $\hat{y}_k$ pour $k = 1, 2, \cdots, m$ où m est le nombre de points.

$$\hat{y}_k = \beta_0 + x_k\, \beta_1 \tag{17}$$

Il n'y a que deux paramètres puisqu'il n'y a qu'une seule caractéristique. Ces deux paramètres sont le coefficient angulaire de la droite (β_1) et son ordonnée à l'origine (β_0). On essaye donc de faire passer une droite au mieux à travers le nuage de points (fig. 1). Les écarts verticaux entre le modèle linéaire et les ordonnées des points sont les termes résiduels ϵ. Dans l'exemple de la figure 1, ces termes résiduels sont distribués selon une loi normale d'écart-type $\sigma = 1$.

Etudions la résolution du problème par l'équation normale. La matrice X est de taille $m \times 2$, donc la matrice $X^T X$ est une matrice $2X2$.

$$X^T X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_m \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix} \tag{18}$$
$$= \begin{pmatrix} m & \sum x_k \\ \sum x_k & \sum x_k^2 \end{pmatrix} \tag{19}$$

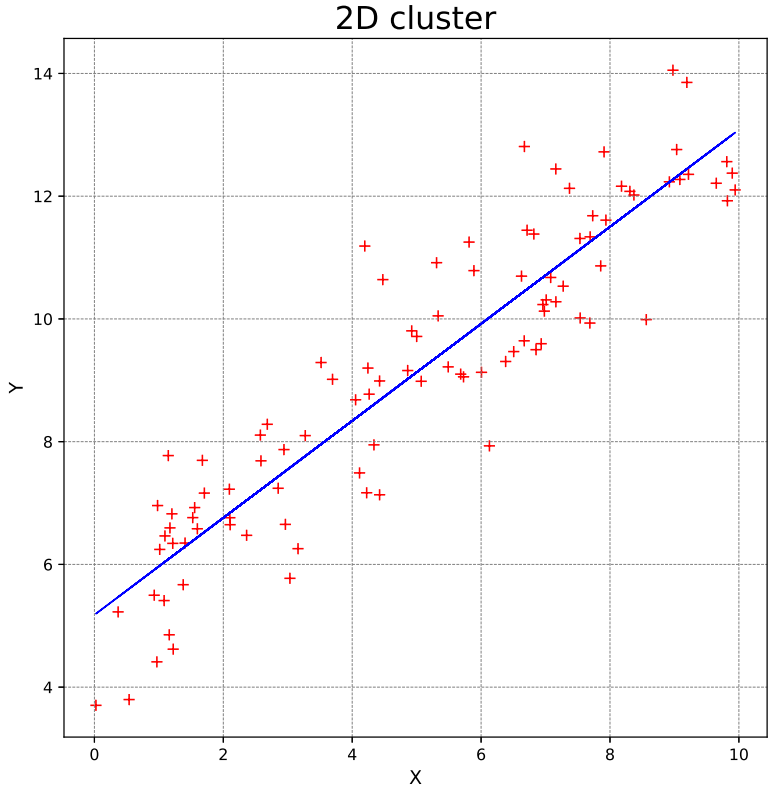

FIGURE 1 – Nuage de points à 2D et modèle linéaire qui minimise la SCER.

Le membre de droite de l'équation normale 8 est

$$X^T \vec{y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_m \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \tag{20}$$

$$= \begin{pmatrix} \sum y_k \\ \sum x_k\, y_k \end{pmatrix} \tag{21}$$

L'équation normale peut donc s'écrire pour un nuage de points à deux dimensions

$$\begin{pmatrix} m & \sum x_k \\ \sum x_k & \sum x_k^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum y_k \\ \sum x_k\, y_k \end{pmatrix}$$

Avant de résoudre ce système linéaire, rappelons que

$$\begin{aligned}
\mathrm{Var}(x) &= \frac{1}{m}\sum_k (x_k - \bar{x})^2 \\
&= \frac{1}{m}\sum_k \left(x_k^2 - 2x_k\bar{x} + \bar{x}^2\right) \\
&= \frac{1}{m}\sum_k x_k^2 - \frac{2\bar{x}}{m}\sum_k x_k + \frac{\bar{x}^2}{m}\sum_k 1 \\
&= \bar{x^2} - 2\bar{x}^2 + \bar{x}^2 \\
&= \bar{x^2} - \bar{x}^2
\end{aligned}$$

et que

$$\begin{aligned}
\mathrm{Cov}(x,y) &= \frac{1}{m}\sum_k (x_k - \bar{x})\,(y_k - \bar{y}) \\
&= \frac{1}{m}\sum_k (x_k\,y_k - x_k\bar{y} - \bar{x}y_k + \bar{x}\bar{y}) \\
&= \frac{1}{m}\sum_k x_k\,y_k - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} \\
&= \frac{1}{m}\sum_k x_k\,y_k - \bar{x}\bar{y} \qquad (22)
\end{aligned}$$

$$(23)$$

Le déterminant de la matrice du système est

$$\begin{aligned}
m\sum x_k^2 - \left(\sum x_k\right)^2 &= m^2\left(\frac{1}{m}\sum x_k^2 - \left(\frac{1}{m}\sum x_k\right)^2\right) \qquad (24)\\
&= m^2(\bar{x^2} - \bar{x}^2) \qquad (25)\\
&= m^2\mathrm{Var}(x) \qquad (26)
\end{aligned}$$

De là, l'inverse de la matrice du système linéaire est

$$\left(X^TX\right)^{-1} = \frac{1}{m^2\mathrm{Var}(x)}\begin{pmatrix}\sum x_k^2 & -\sum x_k \\ -\sum x_k & m\end{pmatrix}$$

On peut alors écrire la solution du système

$$\begin{aligned}
\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} &= \frac{1}{m^2\mathrm{Var}(x)}\begin{pmatrix}\sum x_k^2 & -\sum x_k \\ -\sum x_k & m\end{pmatrix}\begin{pmatrix}\sum y_k \\ \sum x_k\,y_k\end{pmatrix} \qquad (27)\\
&= \frac{1}{m^2\mathrm{Var}(x)}\begin{pmatrix}\sum x_k^2\sum y_k - \sum x_k\sum x_k\,y_k \\ -\sum x_k\sum y_k + m\sum x_k\,y_k\end{pmatrix} \qquad (28)
\end{aligned}$$

D'après 22, $\sum_k x_k\, y_k = m(\mathrm{Cov}(x,y) + \bar{x}\bar{y})$. On a donc pour le coefficient angulaire de la droite de régression

$$\begin{aligned}\beta_1 &= \frac{m^2\left(-\bar{x}\bar{y} + \mathrm{Cov}(x,y) + \bar{x}\bar{y}\right)}{m^2\mathrm{Var}(x)} \\ &= \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}\end{aligned} \tag{29}$$

Pour l'ordonnée à l'origine

$$\begin{aligned}\beta_0 &= \frac{m^2\left(\bar{x^2}\bar{y} - \bar{x}\left(\mathrm{Cov}(x,y) + \bar{x}\bar{y}\right)\right)}{m^2\mathrm{Var}(x)} \\ &= \frac{\bar{y}\left(\bar{x^2} - \bar{x}^2\right) - \bar{x}\,\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} \\ &= \frac{\bar{y}\,\mathrm{Var}(x) - \bar{x}.\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} \\ &= \bar{y} - \frac{\bar{x}\,\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} \\ &= \bar{y} - \bar{x}\beta_1\end{aligned} \tag{30}$$

Avec ces expressions pour les paramètres, nous pouvons déduire une relation simple pour le coefficient de détermination. Partons de la relation 5. Dans le cas d'un nuage de points à deux dimansions, en appliquant 17 nous avons

$$\begin{aligned}SCER &= \sum_{k=1}^{m}\left(y_k - \beta_0 - \beta_1 x_k\right)^2 \\ &= \sum_{k=1}^{m}\left(\left(y_k - \bar{y}\right) - \beta_1\left(x_k - \bar{x}\right)\right)^2 \text{ en utilisant 30} \\ &= \sum_{k=1}^{m}\left(y_k - \bar{y}\right)^2 + \beta_1^2\sum_{k=1}^{m}\left(x_k - \bar{x}\right)^2 - 2\beta_1\sum_{k=1}^{m}\left(y_k - \bar{y}\right)\left(x_k - \bar{x}\right) \\ &= m\mathrm{Var}(y) + \beta_1^2 m\mathrm{Var}(x) - 2\beta_1 m\mathrm{Cov}(x,y) \\ &= m\left(\mathrm{Var}(y) + \frac{\mathrm{Cov}^2(x,y)}{\mathrm{Var}(x)} - 2\frac{\mathrm{Cov}^2(x,y)}{\mathrm{Var}(x)}\right) \text{ en utilisant 29} \\ &= m\left(\mathrm{Var}(y) - \frac{\mathrm{Cov}^2(x,y)}{\mathrm{Var}(x)}\right) \\ &= m\mathrm{Var}(y)\left(1 - \frac{\mathrm{Cov}^2(x,y)}{\mathrm{Var}(x)\mathrm{Var}(y)}\right)\end{aligned} \tag{31}$$

Si on compare cette dernière ligne à 16, cela nous donne

$$R^2 = \frac{\mathrm{Cov}^2(x,y)}{\mathrm{Var}(x)\mathrm{Var}(y)} \tag{32}$$

On définit un coefficient de corrélation linéaire noté ρ par $R^2 = \rho^2$. De ce fait $-1 \leq \rho \leq 1$ et ce coefficient vérifie

$$\rho = \frac{\mathrm{Cov}(x,y)}{\sigma(x)\sigma(y)}$$

6 annexes

6.1 algèbre et calculus

Soient M une matrice de taille m sur n, $\vec{u}$ et $\vec{v}$ des vecteurs de tailles n et m respectivement, nous avons

$$\begin{aligned}
\left(M\vec{u}\right)^T \cdot \vec{v} &= \sum_{i=1}^{m} \left[M\vec{u}\right]_i v_i \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij}u_j v_i \\
&= \sum_{j=1}^{n} u_j \left(\sum_{i=1}^{m} \left[M^T\right]_{ji} v_i\right) \\
&= \sum_{j=1}^{n} u_j \left[M^T\vec{v}\right]_j \\
&= \vec{u}^T \cdot \left(M^T\vec{v}\right)
\end{aligned} \tag{33}$$

Par commutativité du produit scalaire, $\left(M\vec{u}\right)^T \cdot \vec{v} = \vec{v}^T \cdot \left(M\vec{u}\right)$, nous avons donc aussi

$$\vec{v}^T \cdot \left(M\vec{u}\right) = \vec{u}^T \cdot \left(M^T\vec{v}\right) \tag{34}$$

Si dans la relation 33, on considère que $\vec{v} = M\vec{u}$, nous avons

$$\left(M\vec{u}\right)^T \cdot \left(M\vec{u}\right) = \vec{u}^T \cdot \left(M^T M\vec{u}\right) \tag{35}$$

Considérons un vecteur $\vec{u}$ dont les n composantes $u_1, u_2, \cdots$ sont des variables. Notons

$\vec{\nabla}_{\vec{u}} f = \left(\frac{\partial f}{\partial u_1}, \frac{\partial f}{\partial u_2}, \cdots \right)^T$. Nous avons,

$$\begin{aligned}
\vec{\nabla}_{\vec{u}} \left(\vec{u}^T \cdot \vec{v} \right) &= \vec{\nabla}_{\vec{u}} \sum_{j=1}^{n} u_j v_j \\
&= \left(\frac{\partial \sum_{j=1}^{n} u_j v_j}{\partial u_1}, \cdots \right)^T \\
&= (v_1, \cdots)^T \\
&= \vec{v}
\end{aligned} \tag{36}$$

Toujours avec un vecteur $\vec{u}$ dont les n composantes sont des variables et avec M, une matrice carrée $n \times n$. Notons $\vec{e}_i$, les vecteurs de base de notre espace vectoriel.

$$\begin{aligned}
\vec{\nabla}_{\vec{u}} \left(\vec{u}^T \cdot (M\vec{u}) \right) &= \sum_{i=1}^{n} \frac{\partial}{\partial u_i} \left(\sum_{j=1}^{n} u_j \left(\sum_{k=1}^{n} M_{jk} u_k \right) \right) \vec{e}_i \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} M_{jk} \frac{\partial}{\partial u_i} \left(u_j u_k \right) \vec{e}_i \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} M_{jk} \left(\delta_{i,j} u_k + u_j \delta_{i,k} \right) \vec{e}_i \\
&= \sum_{i=1}^{n} \left(\sum_{k=1}^{n} M_{ik} u_k + \sum_{j=1}^{n} M_{ji} u_j \right) \vec{e}_i \\
&= \sum_{i=1}^{n} \left(\sum_{k=1}^{n} M_{ik} u_k + \sum_{j=1}^{n} \left[M^T \right]_{ij} u_j \right) \vec{e}_i \\
&= \sum_{i=1}^{n} \left(\left[M\vec{u} \right]_i + \left[M^T \vec{u} \right]_i \right) \vec{e}_i \\
&= \left(M + M^T \right) \vec{u}
\end{aligned} \tag{37}$$

6.2 Divers

Démontrons que le troisième terme du membre de droite de l'égalité 12 est nul. Il peut se mettre sous la forme vectorielle suivante

$$\sum_{k=1}^{m} \left(y_k - \hat{y}_k \right) \left(\hat{y}_k - \bar{y} \right) = \vec{y}^T \cdot \vec{\hat{y}} - \left\| \vec{\hat{y}} \right\|^2 - \bar{y} \sum_{k=1}^{m} \left(y_k - \hat{y}_k \right)$$

Les deux premiers termes s'annulent entre eux. Ceci est évident suite à l'interprétation géométrique de $\vec{\hat{y}}$ comme la projection de $\vec{y}$ sur un sous-espace vectoriel. Mais, on peut le

montrer par les relations vectorielles.

$$\begin{aligned}
\vec{y}^T \cdot \vec{\hat{y}} - \left\|\vec{\hat{y}}\right\|^2 &= \vec{y}^T \cdot \vec{\hat{y}} - (X\vec{\beta})^T \cdot (X\vec{\beta}) && \text{avec 4} \\
&= \vec{y}^T \cdot \vec{\hat{y}} - \vec{\beta}^T \cdot (X^T X\vec{\beta}) && \text{avec 33} \\
&= \vec{y}^T \cdot \vec{\hat{y}} - \vec{\beta}^T \cdot (X^T\vec{y}) && \text{avec 8} \\
&= \vec{y}^T \cdot \vec{\hat{y}} - (X\vec{\beta})^T \cdot \vec{y} && \text{avec 33 à nouveau} \\
&= \vec{y}^T \cdot \vec{\hat{y}} - \vec{\hat{y}}^T \cdot \vec{y} && \text{avec 4 à nouveau} \\
&= 0 \text{ par commutativité du produit scalaire}
\end{aligned}$$

Il nous reste à montrer que $\sum_{k=1}^{m}(y_k - \hat{y}_k) = 0$. Ce qui revient à montrer que les y_k et les $\hat{y}_k$ ont la même moyenne. Pour cela, considérons la première ligne de la relation vectorielle 10, à savoir $X^T\vec{\hat{y}} = X^T\vec{y}$. Vu que la première ligne de X^T est une ligne de 1. Elle dit en susbstance que

$$\sum_{k=1}^{m}\hat{y}_k = \sum_{k=1}^{m} y_k$$

Ce qui termine notre démonstration.