# Cours sur le coefficient de détermination

**Introduction** Le coefficient de détermination mesure l'adéquation entre un modèle issu d'une régression linéaire simple ou multiple et les données observées (ou les réalisations des variables aléatoires) qui ont permis de l'établir.
Il y a deux façons de l'établir :
Dans le cadre d'une régression linéaire simple, le plus rapide est d'élever au carré le coefficient de corrélation. On le note alors avec une minuscule $r^2$. en revanche, lorsqu'il existe plusieurs séries de variables aléatoires éventuellement explicatives (régression multiple), on le note généralement avec une majuscule $R^2$.

**Les propriétés du coefficient de détermination** Le coefficient de détermination se situe entre 0 (le modèle linéaire ne vaut rien) et 1 (le modèle linéaire est parfait).

La deuxième manière est beaucoup plus riche en implications car elle s'applique aussi bien à une régression simple qu'à une régression multiple.
Soit $y_i$ une valeur prise par la variable que l'on cherche à expliquer. Elle peut être décomposée en deux parties : l'une expliquée par le modèle et l'autre résiduelle, due par exemple à des erreurs de mesure.
La dispersion de l'ensemble des observations se décompose donc en variance expliquée par la régression et en variance résiduelle inexpliquée. La variance totale est la somme des deux.
Le $R^2$ se définit alors comme la proportion de variance expliquée dans la variance totale. Si l'on multiplie ces deux variances par l'effectif n, on peut écrire :

$R^2 = (\frac{SCE}{SCT})$ tel que :
SCE : est la somme des carrés des résidus.
SCE $= \Sigma_i(\hat{y_i} - \bar{y})^2$
SCT : est la somme des carrés totaux.
SCT $= \Sigma_i(y_i - \bar{y})^2$ Par exemple, un coefficient de 0,8 indique que 80 % de la dispersion est expliquée par le modèle de régression.

**Le coefficient de détermination ajusté ($R^2$ ajusté)** Le coefficient de détermination ajusté tient compte du nombre de variables. En effet, le principal défaut du $R^2$ est de croitre avec le nombre de variables explicatives. Or, on sait qu'un excès de variables produit des modèles peu robustes. C'est pourquoi on s'intéresse davantage à cet indicateur qu'au $R^2$. Mais ce n'est pas un véritable carré et il peut même être négatif. Voici deux expressions du $R^2$ ajusté :
$R^2$ ajusté $= R^2 - (\frac{k(1-R^2)}{n-k-1}) = 1 - (\frac{(1-R^2)(n-1)}{n-k-1})$
k étant le nombre de variables explicatives

**Exemple régression multiple et $R^2$ avec le langage R** On suppose qu'on a le tableau suivant :

| y | $x_1$ | $x_2$ |
|---|---|---|
| 4 | 1 | 6 |
| 4 | 3 | 6 |
| 8 | 4 | 7 |
| 12 | 6 | 7 |
| 12 | 6 | 9 |

On remarque que le tableau contient une variable à expliquée y et deux variables explicatives $x_1$ et $x_2$.
Sous le langage R on peut créer les trois vecteurs en utilisant les instructions suivantes :
> y <- c(4,4,8,12,12)
> x1 <- c(1,3,4,6,6)
> x2 <- c(6,6,7,7,9)

Maintenant on met dans la variable n la longueur du tableau
> n <- length(y)
On crée un vecteur "cste" = (1,1,1,1,1)
> cste <- rep(1,n)
On utilise "cbind" pour fusionner les valeurs de cste, $x_1$ et $x_2$
> X <- cbind(cste = cste , x1 = x1 , x2 = x2)
On affiche la matrice générée par
> X
On créé la matrice tr qui est la multiplication de la transposée de X et de X
> tr <- t(X) % * % X (t(X) est la transposée de X)
Afficher tr par
> tr
On calcule le déterminant de tr par
> det(tr)
Pour créer la matrice inverse de t(X) % * % X on utilise
> solve(tr)
> ty <- t(X) % * % y

Le résultat d'estimation est affiché dans la variable hatbeta comme suit :
> hatbeta <- solve(tr) % * % ty

A partir de ces coefficients (ce modèle), on peut calculer les estimateurs $\hat{y}$ puis obtenir les résidus.
> ychap <- X % * % hatbeta
> residus <- y - ychap
afficher les résidus :
> residus
Calculer la SCE par
> SCE <- $sum((ychap - mean(y))^2$
résultat SCE = 58,18182
Calculer la SCT :
> SCT <- $sum((y - mean(y))^2)$
résultat SCT = 64

Calculer $R^2$
> R2 <- SCE/SCT
résultat $R^2$ = 0,909090

La lecture du $R^2$ nous indique que 90 % des variations de y sont expliquées par le modèle.
On calcule maintenant le coefficient de corrélation ajusté (coefficient de détermination ajusté) qui prend en compte le nombre de variables explicatives incluses dans le modèle, il est définit comme suit :
$R^2 = 1 - (\frac{(1-R^2)(n-1)}{n-k-1})$
Avec k le nombre de variables explicatives
> k <- 2
> R2a <- 1 - ((n-1)/(n-k-1)) * (1-R2)
R2a = 0,81
Calculer la variance des erreurs en calculant la SCR appelée aussi l'estimation de la variance des erreurs est calculée comme suit :
SCR = $\Sigma_i \epsilon_i^2 = \Sigma_i (y_i - \hat{y}_i)^2$
La variance des erreurs $\hat{\sigma_\epsilon}^2 = \frac{SCR}{n-k-1}$ > SCR <- $sum((y - ychap)^2)$ résultat SCR = 5,8181 > hatsigma2 <- (SCR/(n-k-1))
Le résultat est 2,909091