Discussion supplémentaire
sur la qualité d'une approximation par les moindres carrés

1. Objectif

Réfléchir une fois pour toute sur les coefficients R > 1 dans le projet de S4.

2. Définitions

2.1. Termes

Valeurs mesurées	$\boldsymbol{y_i}, avec\ 1 \leq i \leq N$, valeurs mesurées en fonction d'une variable indépendante elle aussi mesurée x_i
Valeur moyenne des mesures	$\bar{\boldsymbol{y}} = \frac{1}{N}\sum_{i=1}^{N} y_i$
Valeurs estimées par l'approximation	$\hat{\boldsymbol{y}}_{\boldsymbol{i}} = f(x_i), avec\ 1 \leq i \leq N$, où $f(x_i)$ est la fonction arbitraire d'approximation évaluée au point x_i, fonction qui contient par ailleurs un nombre arbitraire *M* de paramètres $a_1, a_2, \ldots a_M$ approximés par les moindres carrés

2.2. Quantité calculées

Somme des erreurs au carré (SSE) ou *Somme des résidus au carré* (SS_{res}) ou *Somme des écarts inexpliquées*	$SS_{res} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} e_i^2 = SSE$ C'est la somme des écarts entre l'approximation et les mesures. C'est l'erreur que l'algorithme des moindres carrés minimise.
Somme des erreurs de régression au carré (SS_{reg}) ou *Somme des écarts expliquées*	$SS_{reg} = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$
Somme des erreurs totales au carré (SST) ou *Variance des mesures*	$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2 = \sigma^2(y_i)$ C'est la variance des données originales.

3. Coefficient de détermination

On peut entre autres mesurer la qualité d'une approximation par le *coefficient de détermination* $\boldsymbol{R^2}$ dont la définition la plus générale est :

$$\boldsymbol{R^2 = 1 - \frac{SS_{res}}{SST}}$$

On l'interprète soit (a) comme le ratio de l'erreur expliquée par l'approximation sur l'erreur totale présente dans les données, soit (b), si l'on s'en tient littéralement à la formule ci-dessus, à la partie restante quand on enlève le ratio entre l'erreur inexpliquée et l'erreur totale.

Le coefficient de détermination est un nombre inférieur à 1, parce que les résidus d'une approximation ne peuvent être inférieurs à 0. C'est aussi en général une quantité supérieure à 0, parce que les résidus seront en général inférieurs à *SST*. Dans le cas exceptionnel où les résidus seraient supérieurs à la variance des données originales (SS_{res} > SST), le coefficient sera négatif, ce qui indique qu'approximer les données par leur moyenne aurait été préférable à la fonction f(x) proposée, ou en d'autres mots, une droite aurait mieux représenté les données.

3.1. Cas d'une régression linéaire, i.e. $\hat{y}$ = f(x) = mx + b

On peut montrer analytiquement que pour une régression linéaire optimisée par les moindres carrés, la relation suivante tient et la somme des erreurs totales équivaut à la somme des erreurs expliquées et non expliquées :

$$SST = SS_{res} + SS_{reg} \qquad \text{(éq. 2.1)}$$

Ce qui conduit après simplification au coefficient de corrélation de Pearson pour les régressions linéaires par les moindres carrés :

$$R^2 = \frac{SS_{reg}}{SST} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

C'est la formule de l'APP2, qui est incidemment plus proche de la première interprétation (a) ci-dessus. Le coefficient de corrélation de Pearson est un nombre toujours compris entre 0 et 1 si f(x) est une droite, parce que la pire approximation par une droite ne peut pas être pire qu'une approximation par la moyenne, et donc les erreurs de régression ne dépasseront jamais la variance totale.

3.2. Cas général

Il faut utiliser la définition générale si on ne peut démontrer analytiquement ou empiriquement la relation 2.1 ci-dessus!