Cet article s’adresse aux débutants. Mon objectif est de faire en sorte que vous ayez les idées un peu plus claires à propose de R2 et que vous soyez capables d’expliquer par vous-même à une autre personne de quoi il s’agit.

Je suis allez au Sequoia National Park avec un garde forestier. Toute la journée j’ai mesuré des diamètres de troncs et lui me donnait son avis sur l’âge du sequoia en question. Je suis crevé mais content. Il faisait beau, le mec était sympa et s’est révélé être un vrai puit de science. Cerise sur le gâteau j’ai un paquet de données que je vais pouvoir utiliser pour concevoir mon modèle.

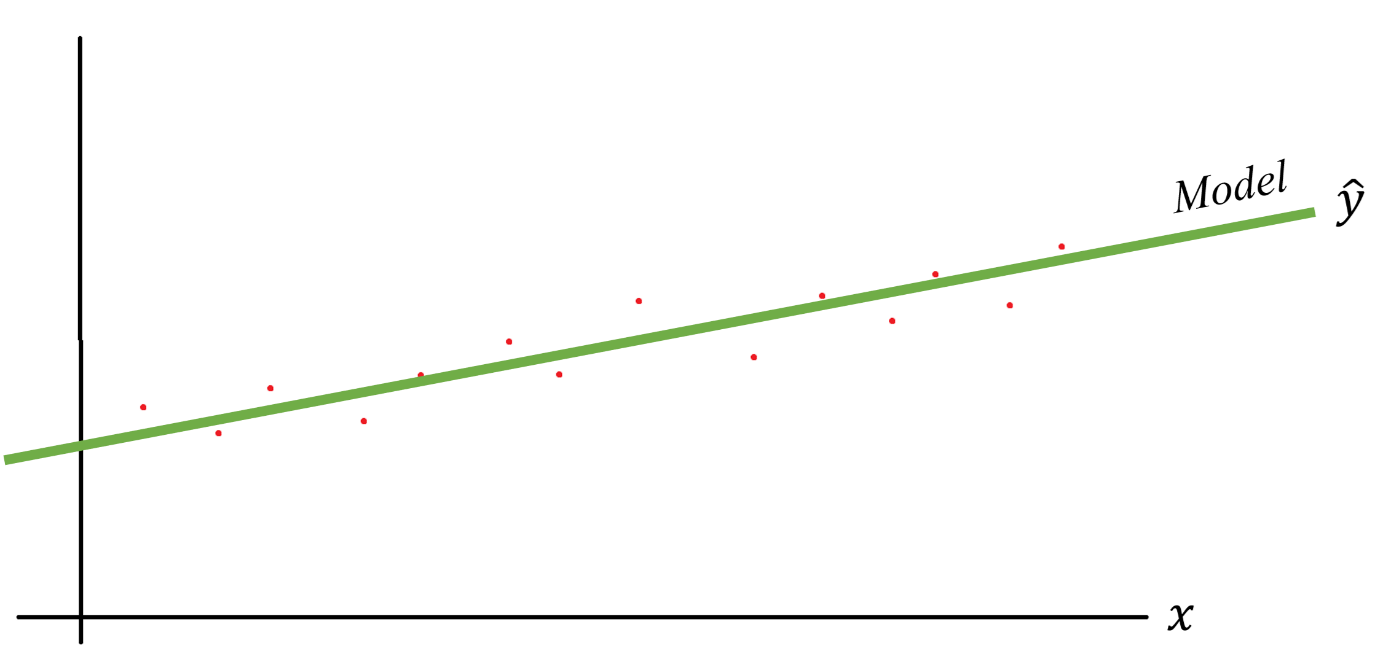
Quand j’affiche l’âge des séquoias en fonction du diamètre de leur troncs voilà ce que je vois :

Une image contenant capture d’écran, ligne, texte, diagramme

Description générée automatiquement

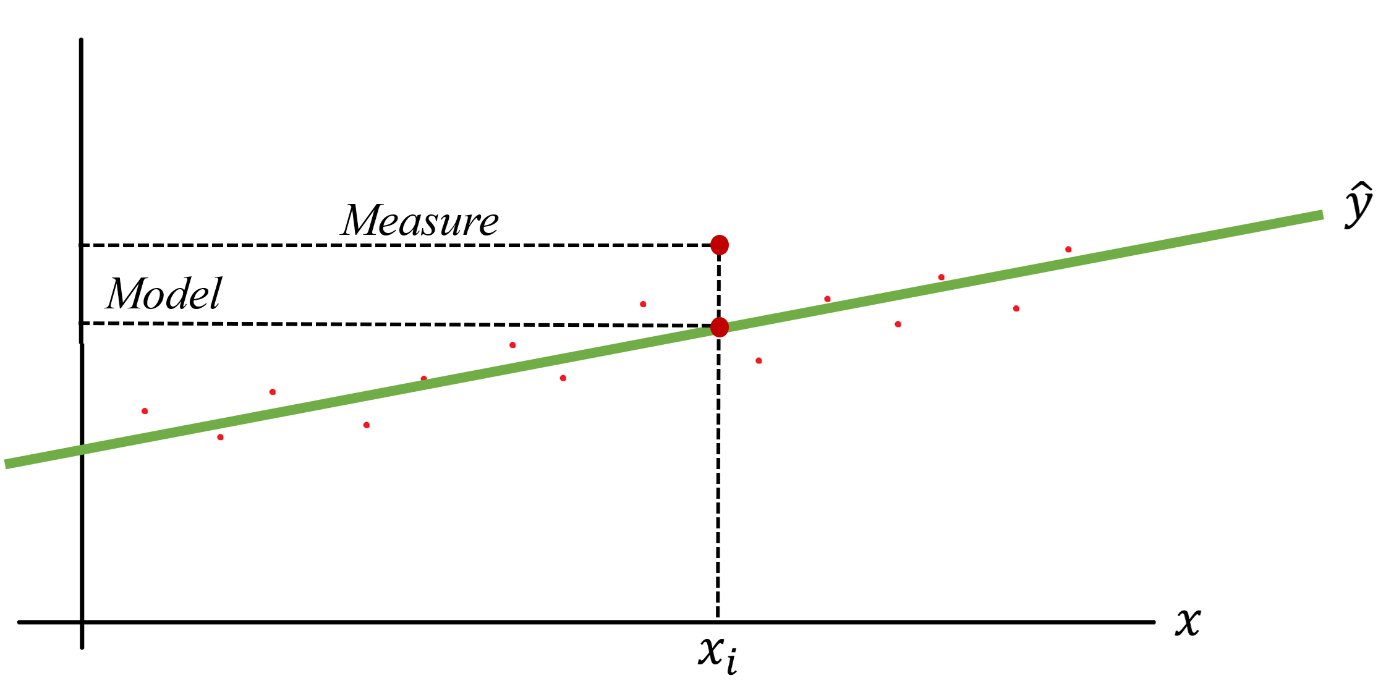
Pour chaque valeur de diamètre (axe horizontal, axe des abscisses) j’ai mis un point rouge à une hauteur (axe vertical, axe des ordonnées) qui correspond à l’âge que m’avait donné le garde forestier.

Vous je ne sais pas, mais mon cerveau n’a qu’une envie… Tracer une droite pour relier tous ces points. Ce que j’ai en tête c’est quelque chose qui ressemble à ça :



Donc, la droite verte c’est mon modèle et on a l’habitude de noter ŷ la valeur prédite par le modèle. Tu notes au passage qu’afin de généraliser, je ne parle plus du diamètre ni de l’âge des séquoias mais d’une variable x et de valeurs y.

OK mais bon, si cela à l’air de fonctionner pour la majorité des points, j’observe qu’il y a toujours un écart entre la mesure faite (les points rouges) et la valeur que propose le modèle. À des fins didactiques, ci-dessous j’ai exagéré l’écart mais voilà en gros ce que j’observe :



Par exemple les diamètres vont de 0 à 11 m (le General Sherman) et j’ai pris la valeur 7 m. Afin de pouvoir, plus tard, généraliser, j’appelle xi cette valeur particulière. Quand je regarde mes notes et le graphique, la valeur annoncée par le garde forestier était 1\_750 ans (voir la ligne en pointillé Measure) alors que pour cette valeur de xi, la droite verte passe par la valeur 1\_500 ans (voir la ligne Model).

Ça c’est pour ce point particulier mais je remarque que c'est pareil pour tous les autres points. Il existe toujours des écarts, en positif ou en négatif, entre les valeurs annoncées par le garde forestier, que je vais considérer comme la vérité et celles prédites par le modèle (la droite verte).

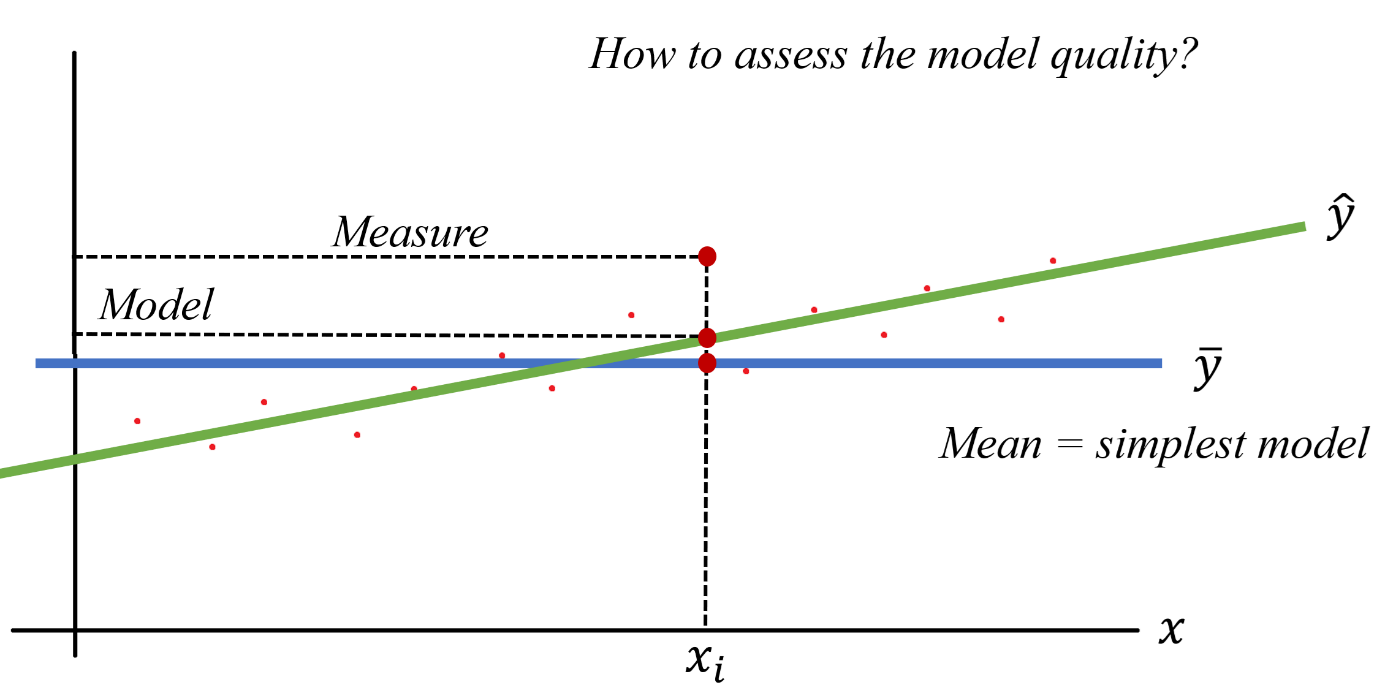
Maintenant, la question à 1 million de dollars est : Heu … Mais que vaut ce modèle ?

Et puis tant qu’on y est… Si demain j’ai un autre modèle, que je trace une droite vert foncé (la couleur est importante) qui a un peu plus de pente. Des deux modèles, comment je peux déterminer celui qui est le plus performant ?

C’est là qu’entre en scène R2. L’idée c’est qu’il nous fournisse, via une seule valeur, un moyen de comparer les performances de différents modèles.

OK… Et comment on fait ? On veut comparer des modèles entre eux. Très bien. Quel serait le modèle le plus simple possible que nous pourrions avoir pour notre histoire de séquoias ? Ce pourrait être un modèle pas très sophistiqué qui répondrait toujours la même chose. Il ne va pas répondre 0 tout le temps car peu de séquoias ont des diamètres de 0m. Il ne va pas, pour les mêmes raisons, répondre 11m (il n’y a qu’un General Sherman). Non, un truc raisonnable qu’un modèle "pas très malin" pourrait répondre c’est la valeur moyenne de toutes les mesures faites.

La situation ressemblerait alors à ça :



J’ai tracé en bleu une droite horizontale à une hauteur ӯ (tu as noté la barre au-dessus du y ?) qui est la valeur moyenne de tous les diamètres mesurés. Il faut bien comprendre que si je dis 2m de diamètre, ce modèle va répondre 1\_250 ans par exemple. De même si j’annonce 8m de diamètre, le modèle va encore répondre 1\_250 ans. Quand je te dis qu’il n’est "pas très malin", il n’est vraiment "pas très malin"…

Si je me place sur une abscisse particulière xi, le modèle simple répond ӯ (1\_250 ans) alors que mon modèle répond ŷi qui est la hauteur (l’ordonnée) à laquelle on se trouve quand, partant de xi on monte à la verticale et qu’on croise la droite verte.

Hmm… Tout à l’heure il y avait un écart entre la vraie valeur et la valeur prédite par le modèle. Maintenant on a un écart entre la valeur prédite par un modèle "pas très malin" et notre modèle. On a gagné quoi dans la bataille ?

LE truc qu’il faut comprendre, c’est que le modèle qui retourne toujours la moyenne va servir d’étalon, de référence. Et dans la suite, ce qu’on va faire, c’est mettre en place une méthode qui va nous permettre de notre modèle avec l’étalon. Si par la suite on compare un autre modèle avec le même étalon, on pourra comparer les 2 modèles entre eux.

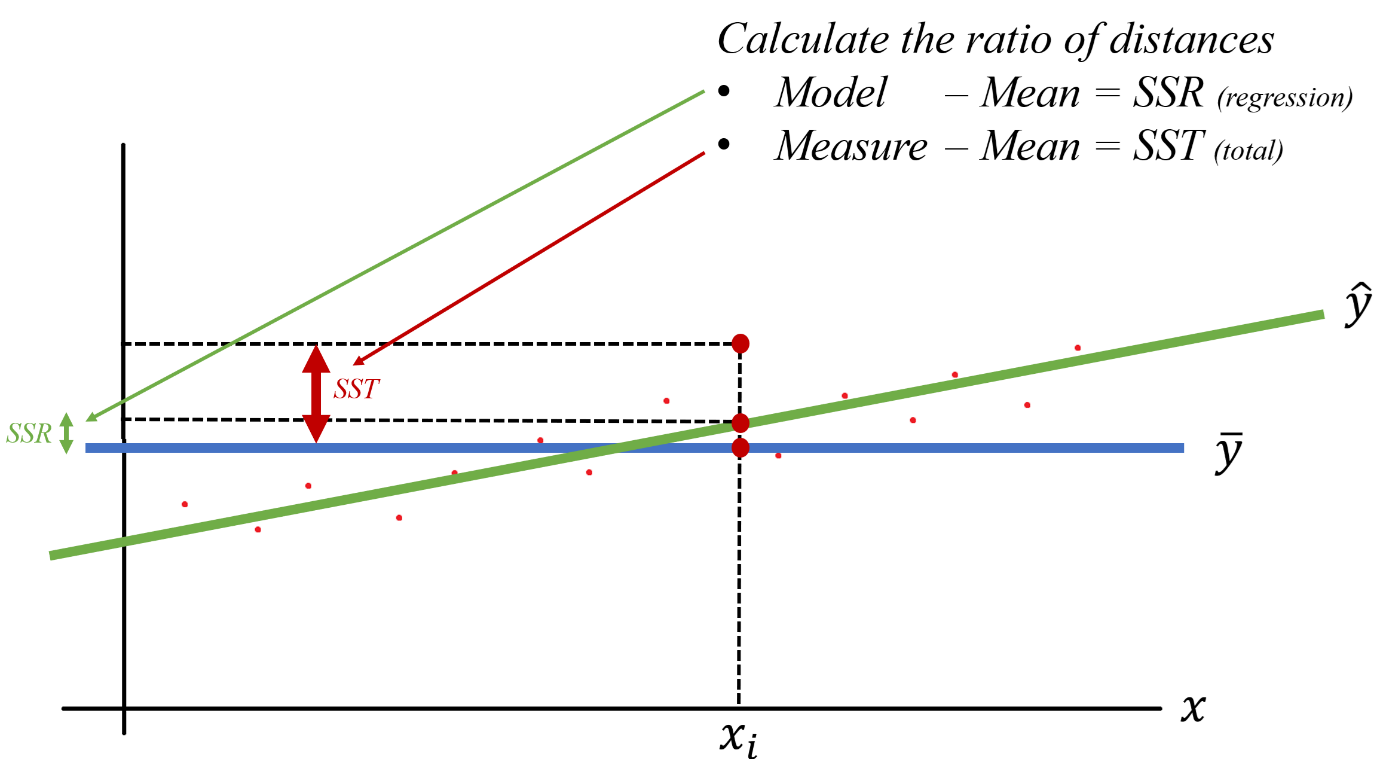
Ok, mais comment on compare 2 grandeurs ? Là, on va faire appel à la loi de Weber qui dit qu’on est plus sensible aux rapports qu’aux différences.

Prenons un exemple pour fixer les idées. Je marche dans la rue. J’observe 2 couples qui marchent devant moi. Dans le premier les 2 personnes mesurent aux environs de 1.75m. La femme est peut-être un peu plus grande mais bon c’est de l’ordre du centimètre. En ce qui me concerne ils ont la même taille. Pour l’autre couple, c’est une autre histoire. Il doit s’agir de Shifang Luo (la championne Chinoise d’haltérophilie qui mesure 1.65m) qui se balade avec Victor Wembanyama (basketteur Français de 2.23m). Là, pour le coup, je "vois" la différence.

Du coup, afin comparer la performance de notre régression linéaire à la mesure réelle, nous calculerons le rapport de leurs écarts respectifs par rapport à la valeur moyenne (notre étalon, notre 0).

Ça va bien se passer tu vas voir et ça va se passer comme ça :

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*



Pas de panique. Notre modèle est une régression avec un "R". Donc ci-dessus, SSR c’est la distance entre la valeur prédite par le modèle de base (je vais arrêter de dire "pas trop malin" c’est un coup à ce qu’il se vexe…) et notre modèle. C’est la différence entre les droites bleue et verte.

Ensuite, on va appeler SST, avec un "T" comme dans Totale, la distance, qu’il y a entre la valeur vraie annoncée par le garde forestier et la moyenne qu’annonce le modèle de base.

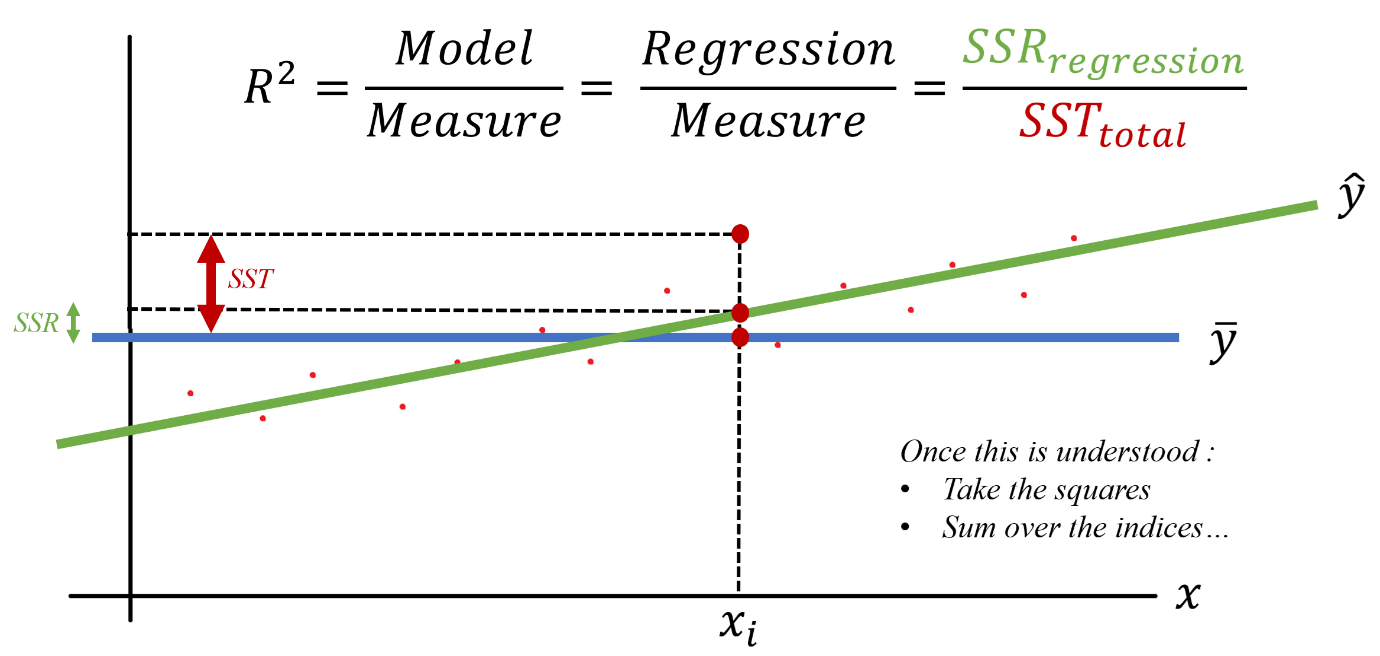
Attends, attends…C’est quoi ces "SS" qui arrivent de nulle part ? Là pour le coup je te demande de me faire confiance. On va les justifier dans 2 paragraphes. Trust in me, just in me…

OK mais tu avais dit qu’on allait comparer des valeurs. Du coup, avec tes termes je m’attendais à ce que l’on compare SST (la différence entre la valeur annoncée par le modèle de base et la valeur vraie) et un truc qui n’a pas de nom sur ton schéma mais qui serait la différence entre la courbe verte et la valeur vraie annoncée par le garde forestier. C’est une très bonne remarque. En fait il y a 2 manières de présenter les choses et j’ai fait un choix. Là aussi soit patient, j’en parle à la fin de l’article.

Du coup on a presque terminé et à ce stade on va dire que

R2 = TODO

La situation ressemble donc à ça.



Oui bien sûr ce n’est pas tout à fait terminé et on peut même dire que c’est complètement faux. En revanche il faut garder en tête que R2 c’est la comparaison (le rapport) de la modélisation et ce qui se passe, en moyenne, dans la vraie vie. Si le rapport vaut 1, notre modèle est formidable, il colle à la vraie vie. Si le R2 vaut 0 cela veut dire que notre modèle n’est pas plus performant que le modèle de base.

R2 = TODO = Modèle/Mesure

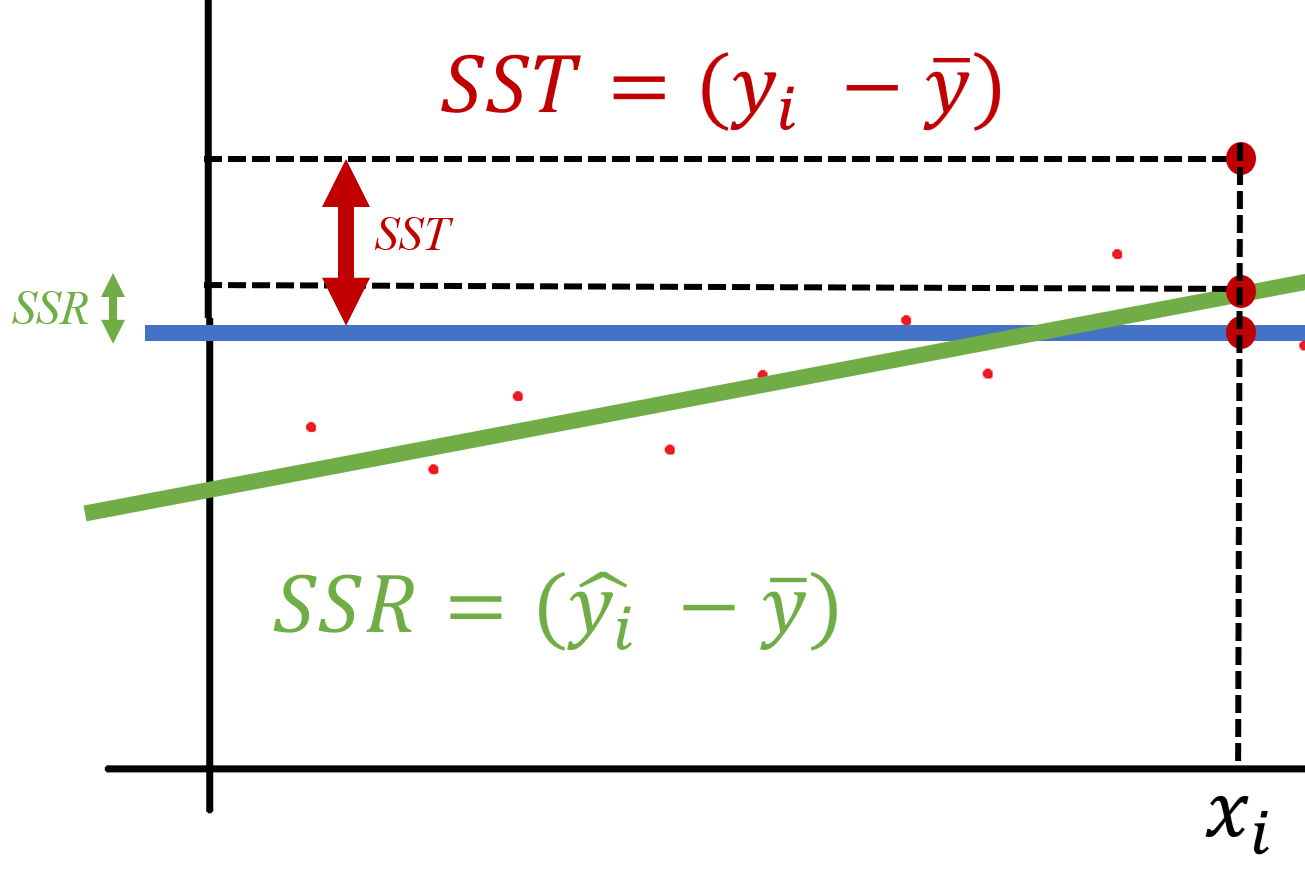
Attends, attends… Tu peux revenir sur la dernière phrase ? Si R2 vaut 0 le modèle est nul et il faut passer à autre chose. Non ? Non, pas tout à fait. Si R2 vaut 0 cela veut dire que le numérateur vaut 0. Il n’y a donc pas de différence entre la prédiction de notre modèle et la prédiction du modèle de base. Ci-dessus, pour être dans ce cas-là, il "suffit" de faire descendre un peu la droite verte pour quelle croise la droite bleue quand x vaut xi. Donc quand R2 vaut 1 le modèle colle parfaitement à la réalité. Quand R2 vaut 0 il n’est pas nul. Il n’est juste pas meilleur que le modèle de base qui répond toujours "moyenne", "moyenne", "moyenne" quel que soit le diamètre qu’on lui propose.

Cela dit, c’est comme pour les contrats d’assurance et le diable est dans les détails. Le R2 qu’on a calculé n’est valable que pour la valeur xi. Ce dont on a besoin, c’est un chiffre qui synthétise la performance globale de notre modèle. Autrement dit un R2 qui couvre tous les xi possibles.

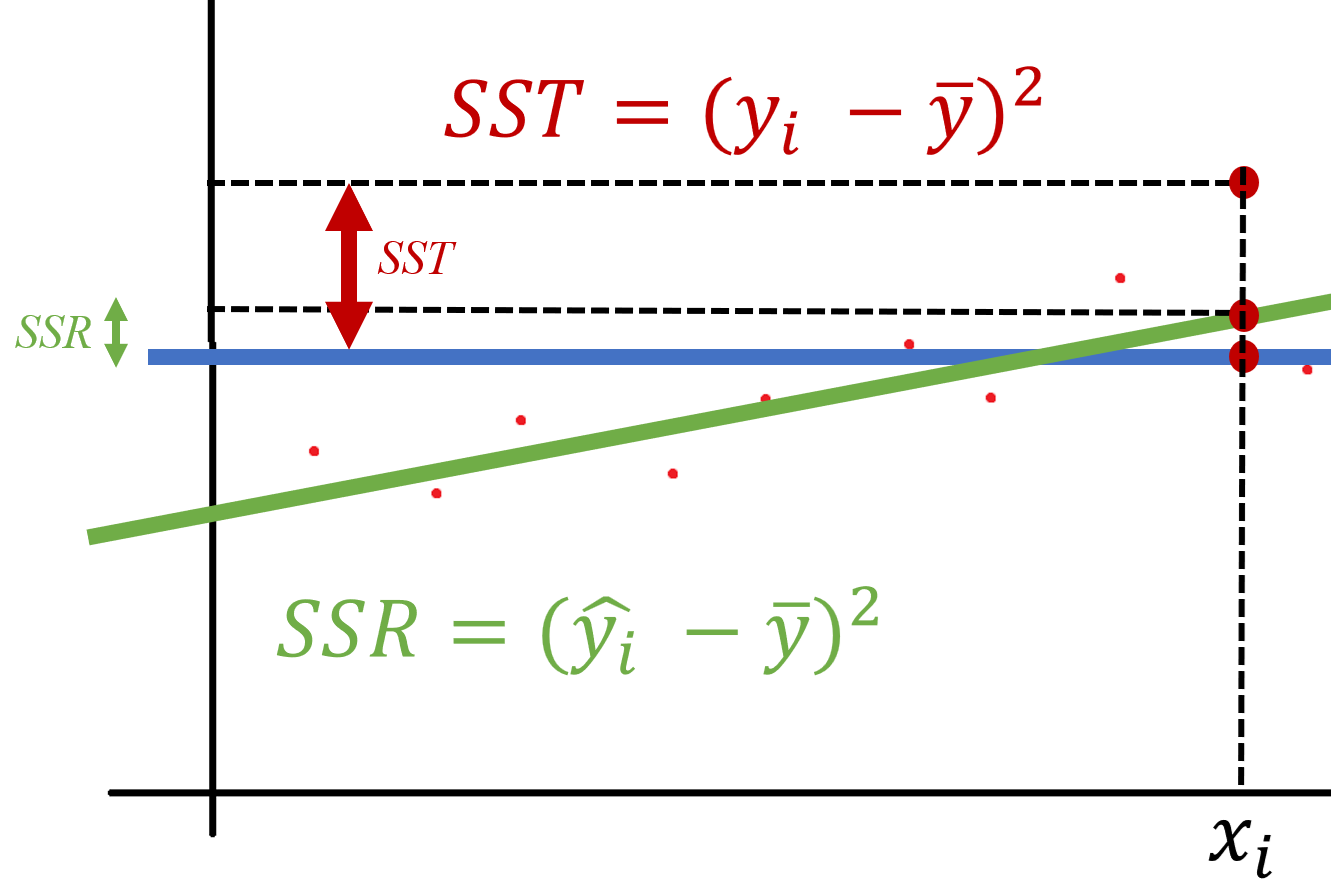
Du coup, on a qu’à faire le ratio des sommes de toutes les différences. Oui mais non. Oui c’est une très bonne idée que de faire la somme mais non, on ne peut pas utiliser une somme de différences. En effet si en x3 la différence vaut -4 et si en x42 la différence vaut +4, les deux erreurs vont se compenser alors qu’en fait elle se cumulent. Du coup on va utiliser la technique de Gauss et on va faire la somme des différences au carré. Comme ça on a plus que des nombres positifs que ne peuvent pas se compenser.

Heu… Somme des différences au carré. Somme avec un "S" et carré avec un "S" aussi comme dans sqared.

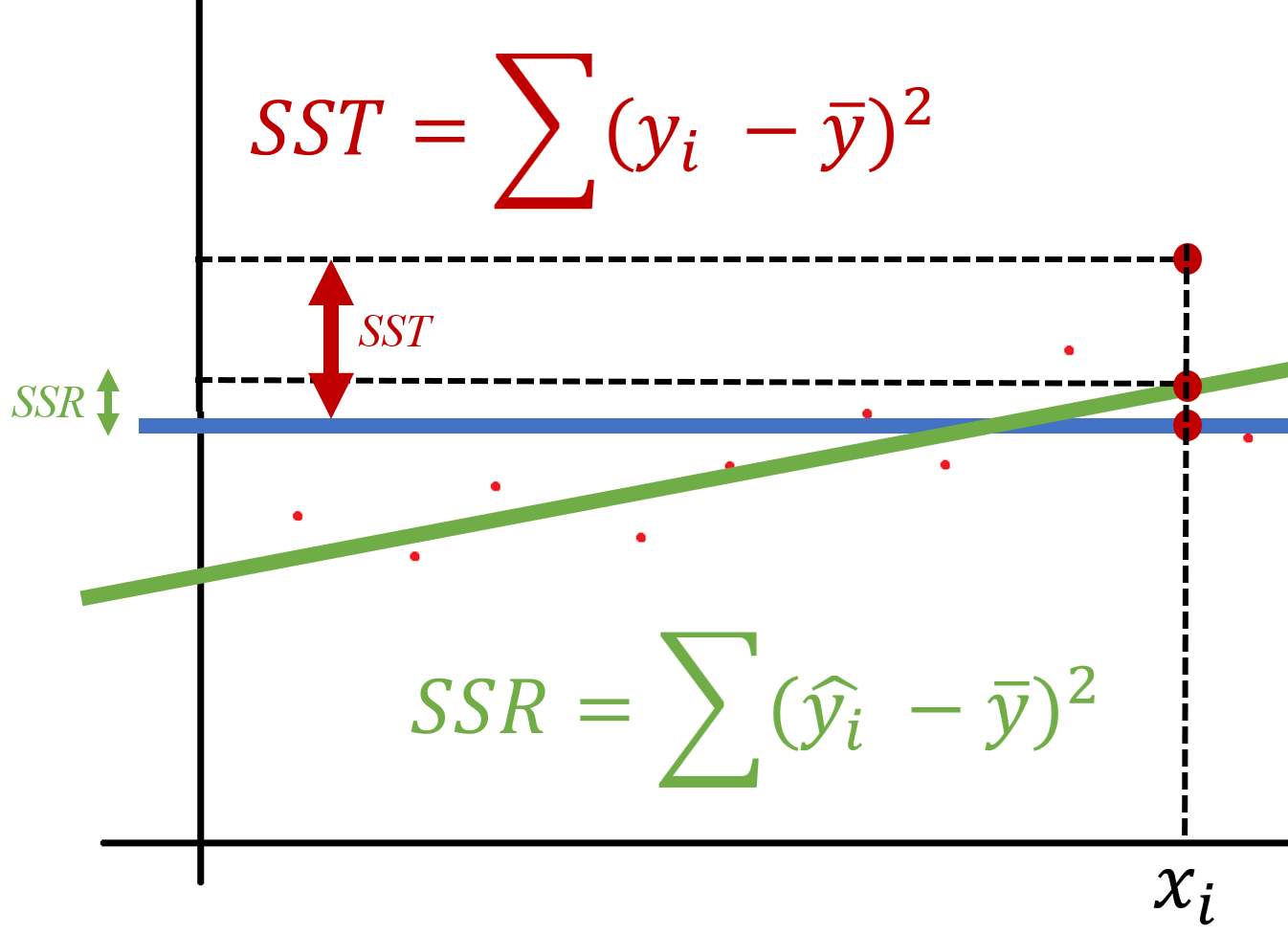
Bref, ça va se passer comme ça :



Pour chaque indice i on calcule SSRi et SSTi.

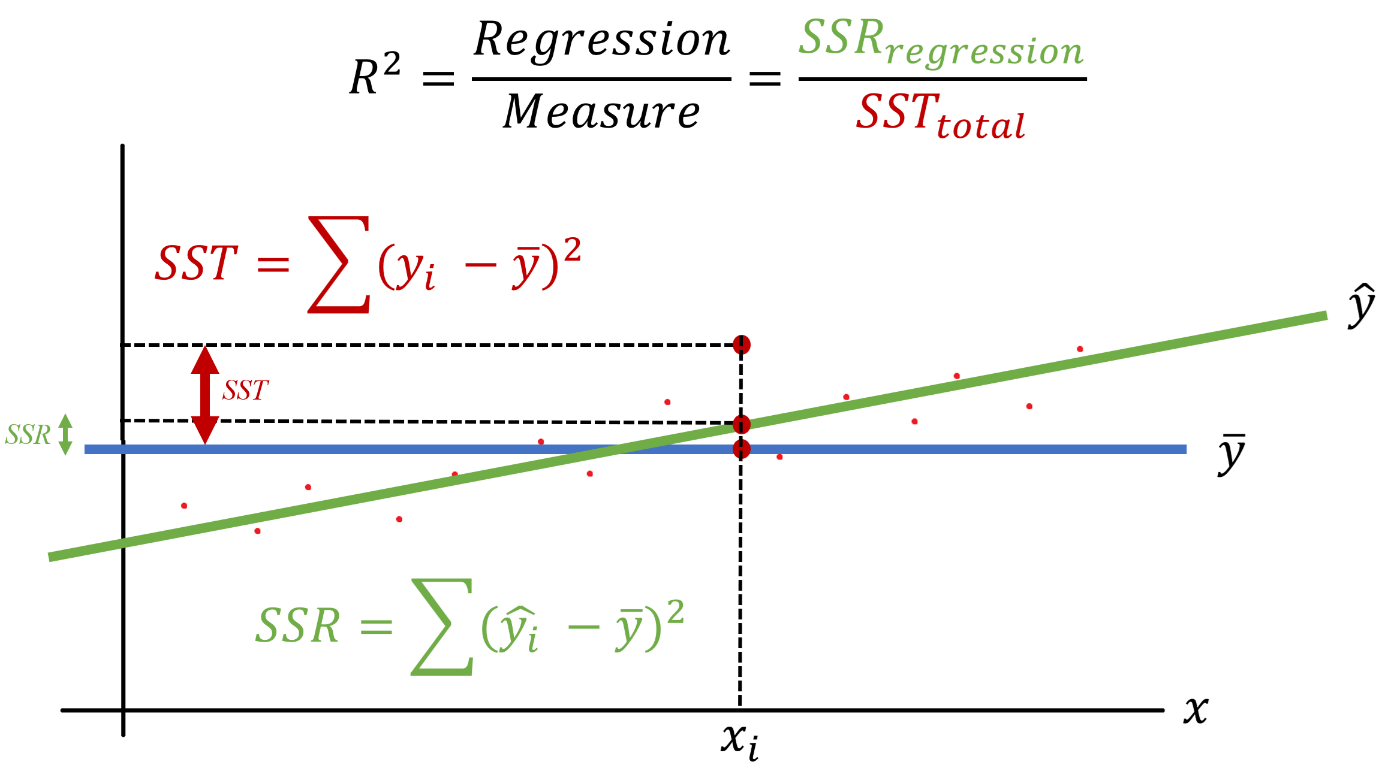


On élève ensuite au carré les 2 valeurs calculées.



On ajoute toutes les valeurs ainsi calculées et à la fin on a les bonnes et vraies valeurs de SST, SSR et R2.

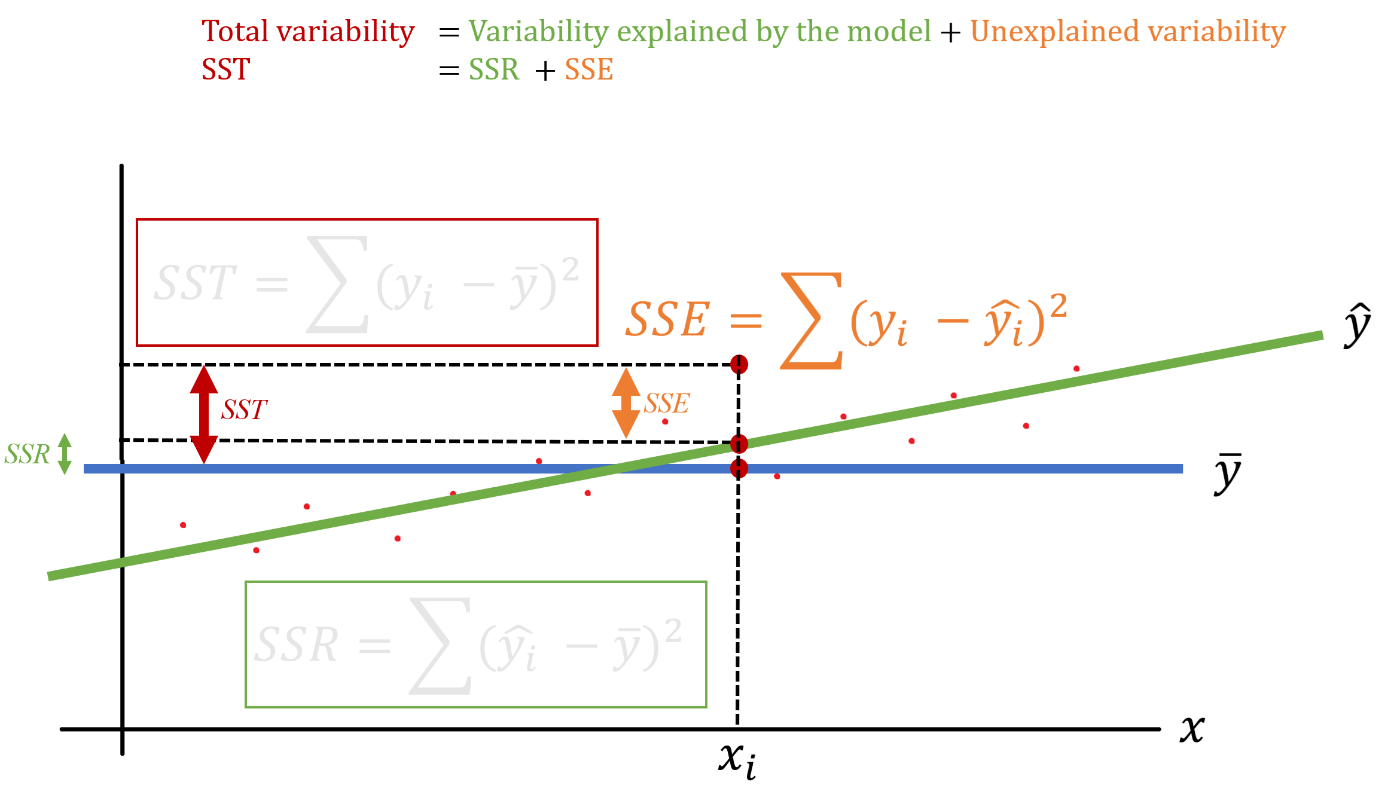
Finalement la situation est donc la suivante :



R2 synthétise donc en une seule valeur la capacité du modèle à expliquer la variabilité des mesures par rapport à leur moyenne. Encore une fois, si le R2 du modèle vaut 1 cela signifie que le modèle explique 100% de la variabilité. Si le R2 du modèle vaut 0, cela signifie que le modèle n’est pas plus performant que le modèle de base qui prédit toujours "moyenne", "moyenne", "moyenne" (ça me fait toujours penser au "spam", "spam","spam" des Monty Python)

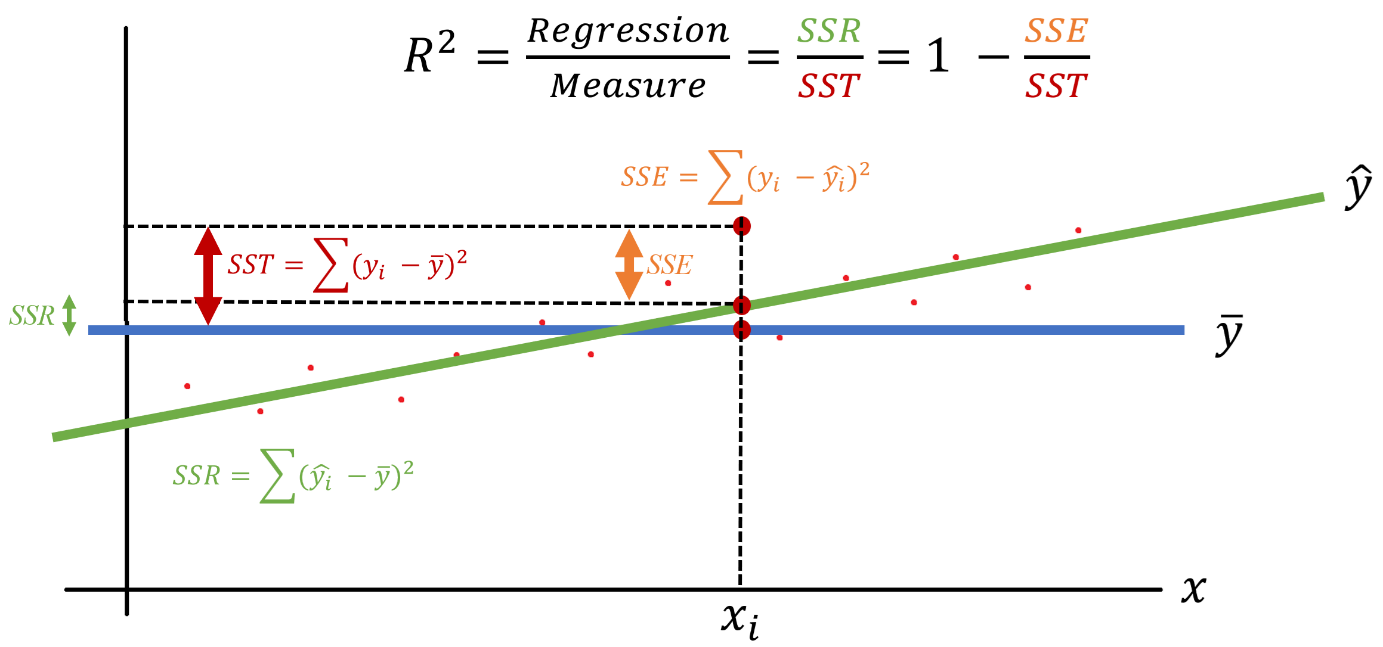
Un dernier point avant de récapituler tout ça. On a dit que R2 c'est le rapport de la prédiction sur ce qui se passe en moyenne dans la vraie vie (SSR sur SST). Cela étant posé, si on remarque que la variabilité totale c'est la variabilité expliquée par le modèle plus la variabilité non expliquée, on va pouvoir introduire le terme dont on n'a pas le droit de prononcer le nom, Lord Voldemort, non, SSE.

On peut illustrer la situation de la façon suivante :



Il faut juste remarquer qu'on a écrit : SST = SSR + SEE ce qui en Français veut bien dire que la variabilité totale (SST) c'est celle expliquée par le modèle (SSR) additionnée à la variabilité non expliquée par le modèle(SSE).

Du coup peut écrire R2 de deux façons différentes :



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Notes de synthèse

1. On a travaillé en 2D car c'est beaucoup plus facile pour les illustrations mais bien sûr tout ce qui a été dit précédemment est valable dans les dimensions supérieures. Imaginons qu'outre le diamètre j'ai mesuré la hauteur des séquoias. Avec x le diamètre et y la hauteur on détermine z l'âge de l'arbre. Du coup notre droite verte devient un plan vert. Même chose si je tiens compte en plus de "je ne sais pas quoi" (j'y connais rien en séquoia en fait). J'ai 3 features x, y et z. Le plan devient un cube de régression linéaire et là, à part dessiner son ombre… Je ne vois pas trop quoi faire.
2. R2 synthétise en une unique valeur la capacité du modèle à expliquer la variabilité des mesures par rapport à leur moyenne.
3. Si le R2 du modèle vaut 1 cela signifie que le modèle explique 100% de la variabilité de mesures.
4. Si le R2 du modèle vaut 0, cela signifie que le modèle n’explique pas plus la variabilité des mesures que le modèle de base qui répond toujours "moyenne", "moyenne", "moyenne"
5. SST = SSR + SSE. En anglais dans le texte cela veut dire que la variabilité totale (SST) c'est celle expliquée par le modèle (SSR) additionnée à la variabilité non expliquée par le modèle(SSE)
6. R2 = TODO = Regression/Mesure = SSR/SST = 1 – SSE/SST
   * SSR = Distance between the mean and the predicted value
   * SST = Distance between the mean and the actual value
   * SSE = Distance between the actual values and the predicted value

TODO

* Pensez faire des recherches sur adjusted R2
* Pensez à lire cet article : [What’s Wrong With R-Squared (And How to Fix It) | by Samuele Mazzanti | Towards Data Science](https://medium.com/towards-data-science/whats-wrong-with-r-squared-and-how-to-fix-it-7362c5f26c53)