REGRESSION LINEAIRE

17 février 2022

Barbara Fraenckel SIMPLON.CO

PROGRAMME

- Pour bien démarrer : ma boite à outil de statistiques
- Application à la régression linéaire simple
- Application à la régression linéaire multiple
- Pour aller plus loin

Pour bien démarrer

Définition

Vocabulaire statistique

Les qualités d’un estimateur

les tests d’hypothèses

Seuil de significativité : cas de la p-value

Pour bien démarrer

définition

on cherche à traduire une problématique concrète par une modélisation mathématique.

C'est-à-dire que l'on va modéliser les valeurs d'une variable, notée Y, en fonction d'une ou plusieurs autres variables notées X_i.

Le modèle, défini par des équations, va servir soit à décrire les phénomènes, souvent dans une optique de causalité, soit à prédire de nouvelles valeurs.

Les différentes appellations possibles :

dépendante vs indépendante

Exogène vs endogène

Régressé vs régresseur

Expliquée vs explicative

Pour bien démarrer

Vocabulaire statistique

Statistiques descriptives

- ✓ **Caractéristiques de position** : moyenne, médiane
- ✓ **Caractéristique de forme** : coefficient de symétrie (skewness) , coef d'aplatissement (kurtosis)
- ✓ **Caractéristiques de dispersion** : variance, écart type, étendue, quartiles (boites à moustache)

→ Sert a repérer les valeurs atypiques , les valeurs aux extrêmes, visualiser les tendances et dans notre cas à voir si la régression linéaire est envisageable

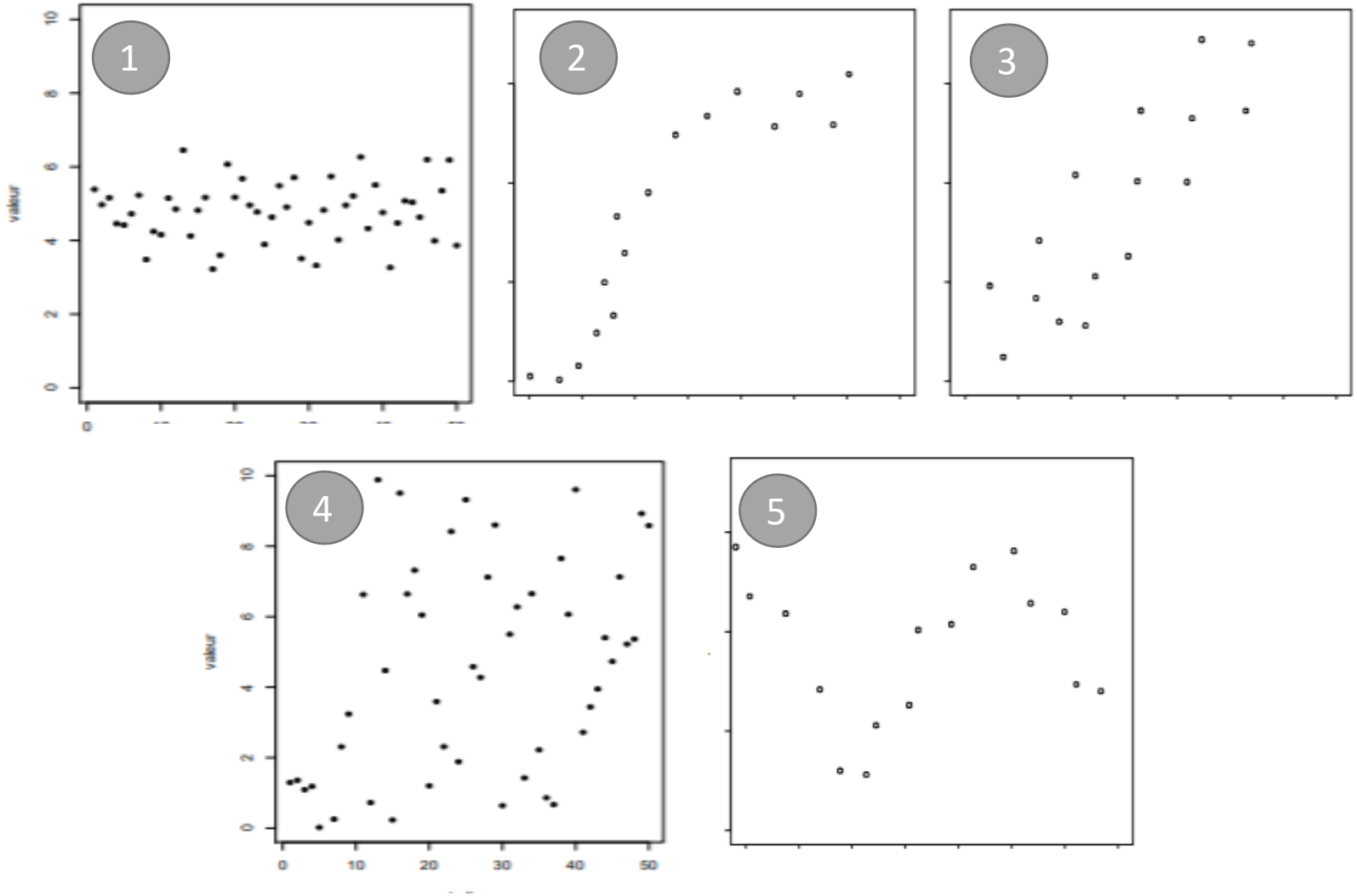
statistiques inférentielles

- ✓ Définition des **estimateurs**
- ✓ évaluation de la qualité des estimateurs : la convergence et la précision (biais , variance)
- ✓ **Les Tests d'hypothèse**
- ✓ seuil de **significativité**

→ Permet de prédire et estimer les caractéristiques et le comportement d'une **population** à partir d'un **échantillon** aléatoire de données sur cette population.

Pour bien démarrer

L’intérêt de la visualisation



Pour bien démarrer

Les qualités d'un estimateur

« En mathématiques, un estimateur $\hat{\theta}$ est une statistique permettant d'évaluer un paramètre inconnu θ relatif à une loi de probabilité. Il peut par exemple servir à estimer certaines caractéristiques d'une population totale à partir de données obtenues sur un échantillon comme lors d'un sondage. » Wikipédia

Les qualités requises :

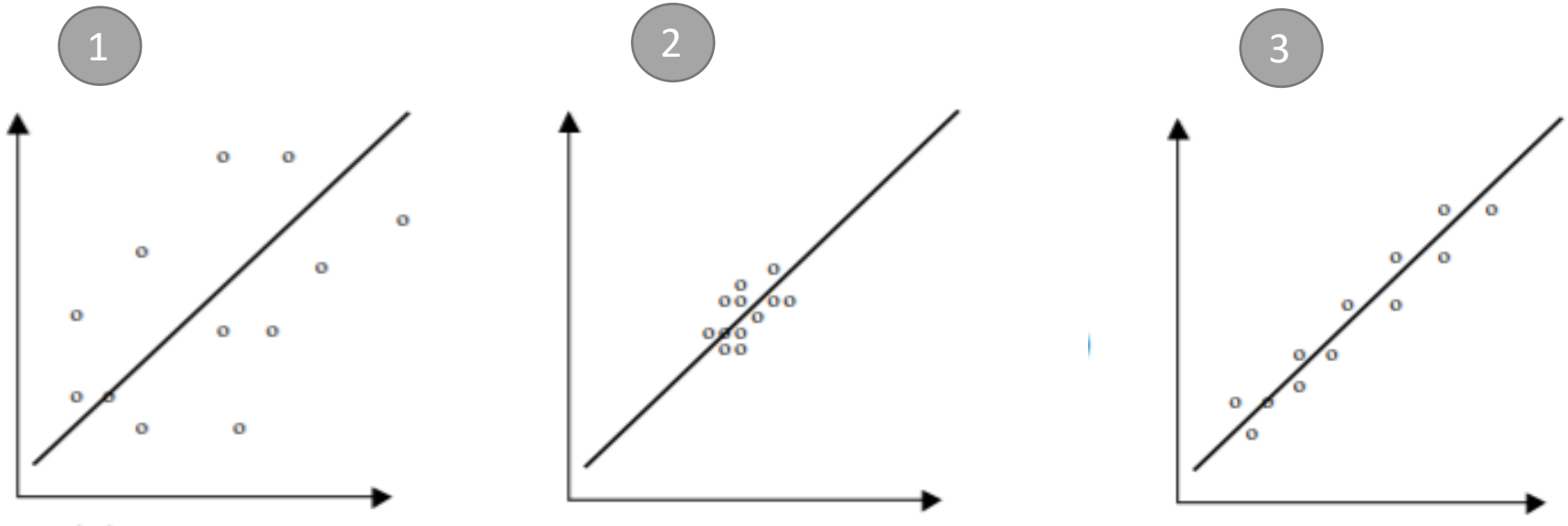
- **Convergence**
- **précision**: se mesure à l'aide du biais et de la variance

Pour bien démarrer

Les qualités d'un estimateur

Les estimateurs sont d'autant plus précis que:

- **La variance de l'erreur est faible**
- **La dispersion des X est forte**



Pour bien démarrer

Veille technique

Expliciter le dilemme du biais/variance d'un estimateur

A quoi est –il du ? Comment le contrer ? Quelles sont les conséquences sur un jeu de donnée en ML ?

Voici quelques liens pour vous aider :

- https://fr.wikipedia.org/wiki/Dilemme_biais-variance
- https://www.datacorner.fr/biais-variance/
- https://openclassrooms.com/fr/courses/4011851-initiez-vous-au-machine-learning/4092326-trouvez-le-bon-compromis-entre-biais-et-variance

Pour bien démarrer

les tests d'hypothèses

Un test d'hypothèse consiste à rejeter ou à accepter une hypothèse statistique, appelée hypothèse nulle **H0**, en fonction d'un jeu de données

Tests d'homogénéité

→ on compare deux échantillons entre eux.

H_0 = les 2 échantillons sont homogènes (semblables)

Tests de conformité

→ on veut déterminer si un échantillon suit une loi statistique connue.

H_0 = vrai (l'échantillon suit la loi).

Pour bien démarrer

les tests d'hypothèses

- **Le test de Student ou Student-Fisher :** sert à la comparaison d'une moyenne observée avec une valeur attendue.
- **Le test de Fisher** : sert à la comparaison de deux variances observées.
- **L'Analyse de la variance ou ANOVA** : sert à comparer plusieurs moyennes observées entre-elles. Il se base sur une décomposition de la variance en une partie " explicable " et une partie " erreur ", supposée distribuée selon la loi normale.
- **Le test de Khi-2 :** sert à la comparaison de deux distributions observées.
- **Le test de Kolmogorov-Smirnov** (comme le test de Khi-2) est un test d'adéquation entre des échantillons observés et une distribution de probabilité. Il compare la fonction de répartition observée et la fonction de répartition attendue. Il est particulièrement utile pour les variables aléatoires continues.

Pour bien démarrer

Le seuil de significativité : cas de la p-value

- On part de l'hypothèse qu'H0 est vrai
- On s'accorde une probabilité, un risque acceptable, de rejeter cette hypothèse nulle alors qu'elle est en fait vraie.
- Ce risque d'erreur α est souvent fixé à 5%.
- On calcule ce risque-là sur cette étude en particulier , c'est ce qu'on appelle la p-value

on dira que l'étude est statistiquement significative si $p < \alpha$.

*La valeur p n'est PAS la probabilité que l'hypothèse nulle soit vraie. La valeur p est en fait la probabilité d'obtenir un échantillon comme le nôtre, ou plus extrême que le nôtre **SI** l'hypothèse nulle est vraie.*
Donc plus la valeur de p est petite, plus la probabilité de faire une erreur en rejetant l'hypothèse nulle est faible.

https://www.mathbootcamps.com/what-is-a-p-value/

Application à la régression linéaire simple

Veille technique

- Test de Student → Groupe 1 et 2
- Test F et ANOVA → Groupe 3 et 4
- Critère d'Akaïke → Groupe 5 et 6
- Test du Ki2 → groupe 7

En quoi ça consiste ?

Quand l'appliquer ?

Comment il fonctionne ?

Comment l'implémenter en python ?

Vous disposez de 45 min pour vous concerter et organiser une restitution de 3 slides

Vous aurez 5 min pour nous le présenter

Faisons le point

Nous avons appris :

- **La différence entre statistique descriptive et inférentielle**
- **Ce qu’est un estimateur**
- **Quelles sont les qualités d’un bon estimateur**
- **Ce qu’est la p-value**
- **Les différents tests statistiques existants**

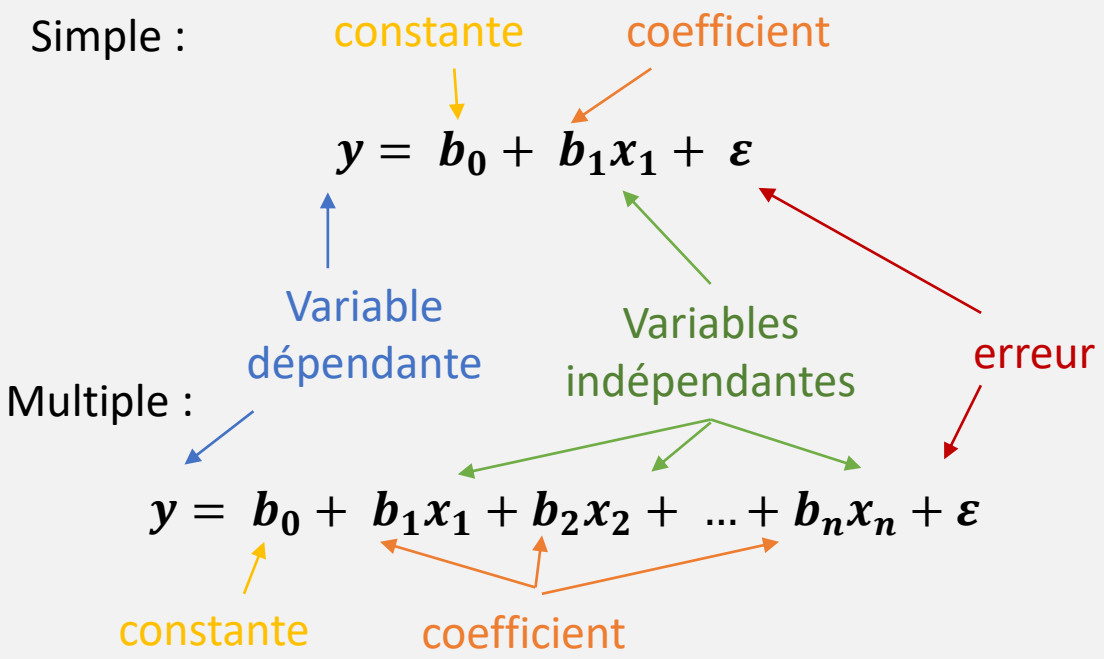
Application à la régression linéaire simple

Les caractéristiques d’une régression
la régression linéaire simple
Méthodes des moindres carrés
R-squared
Adjusted R-squared

Pour bien démarrer:

Les caractéristiques de la régression

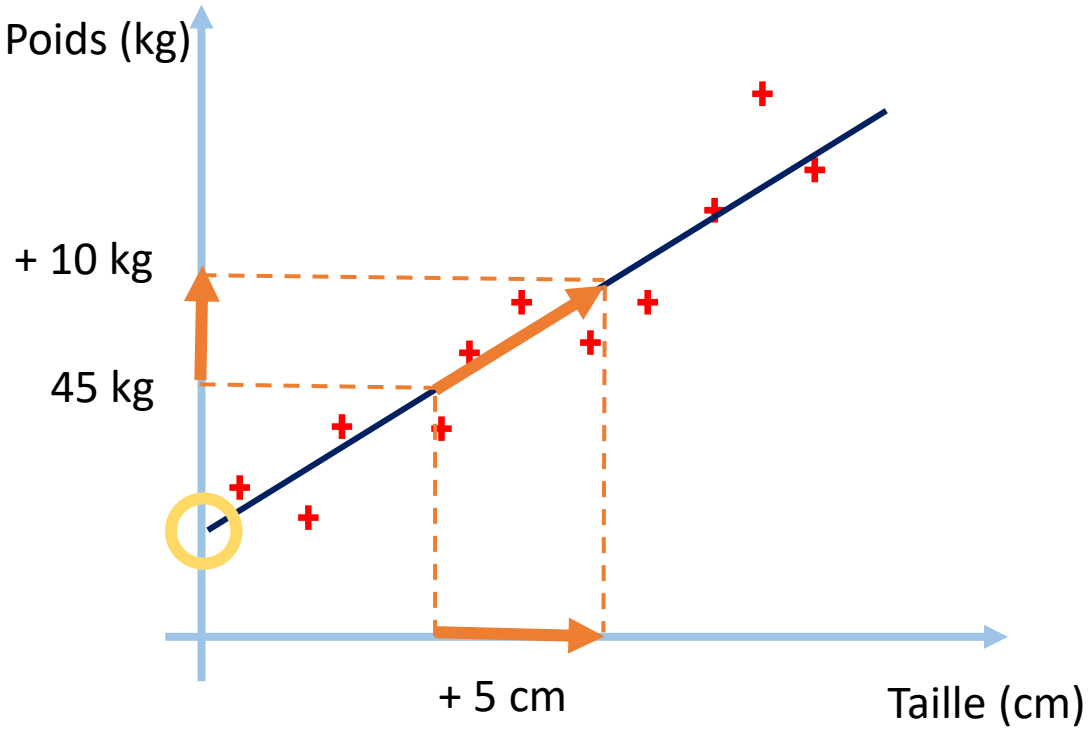
Régression linéaire



Variables quantitatives et non qualitatives

Application à la régression linéaire simple

La régression linéaire simple : Définition



$$y = b_0 + b_1 x$$

$$y = b_0 + b_1 * x_1 + \varepsilon$$

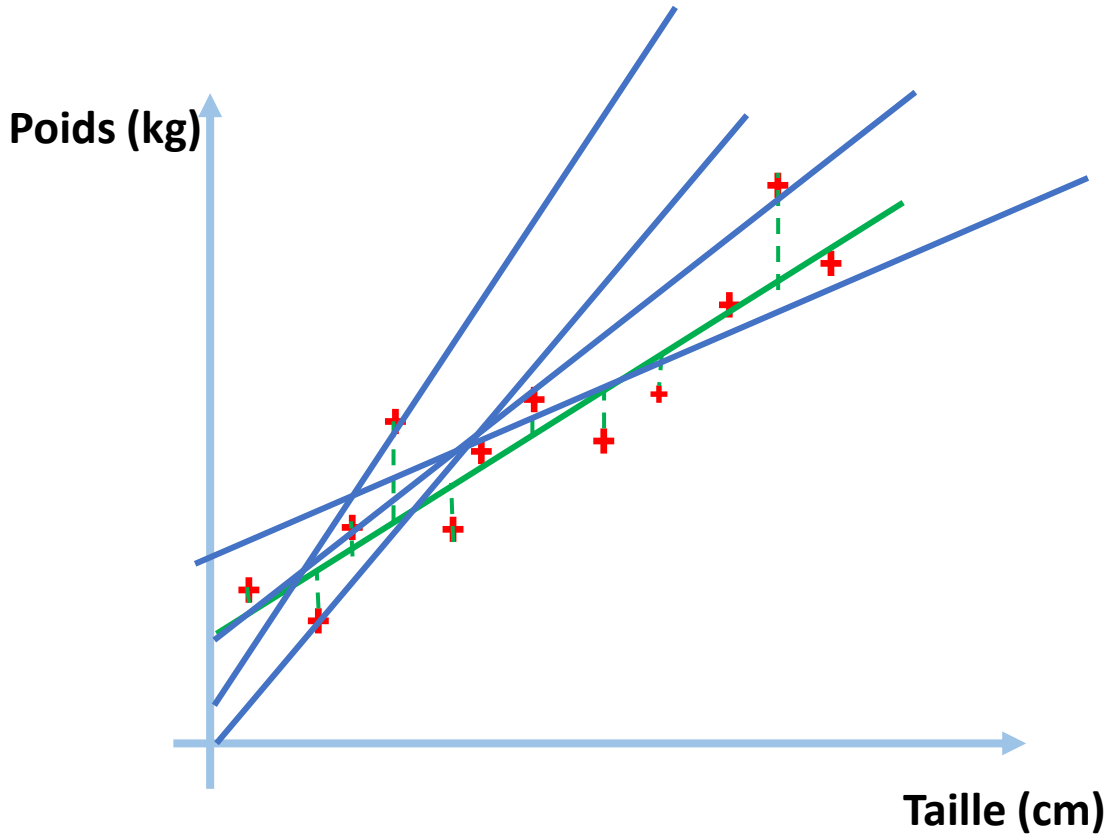
Variable dépendante

Variable indépendante

Poids = b_0 + b_1 * taille

Application à la régression linéaire simple

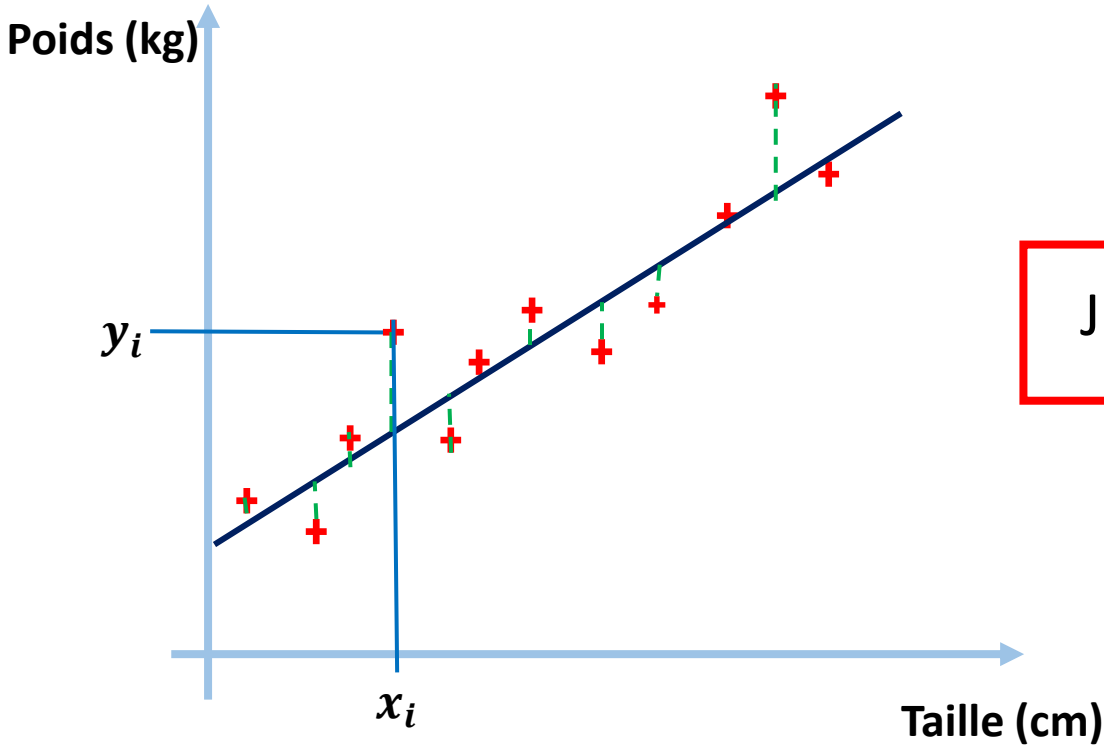
Objectif : réduire la fonction de coût associée



On cherche à trouver un modèle de régression à appliquer : au départ, les valeurs des coefs sont aléatoires. On mesure les erreurs produites. C'est ce qui va déterminer notre fonction de coût. L'objectif sera de minimiser cette fonction pour que notre modèle soit le plus proche de nos données.

Application à la régression linéaire simple

Objectif : réduire la fonction de coût associée



On cherche à minimiser:

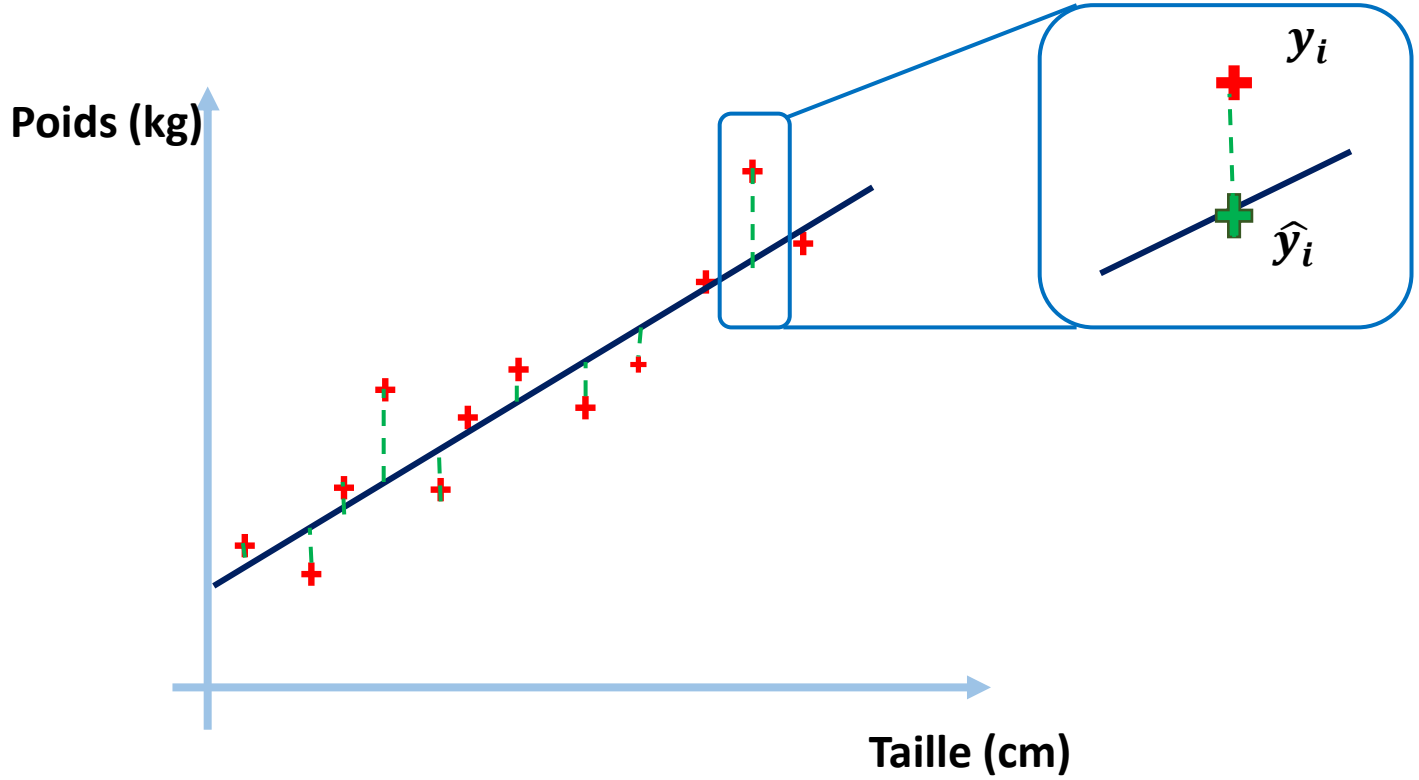
$$J(a,b) = \frac{1}{2m}\sum(f(x^i) - y^i)^2$$

Application à la régression linéaire simple

Le coefficient de détermination R^2

Le coef de détermination évalue à quel point la droite du modèle est meilleure que celle obtenue par la moyenne des observations .

→ on utilise la méthode des moindres carrés ordinaires

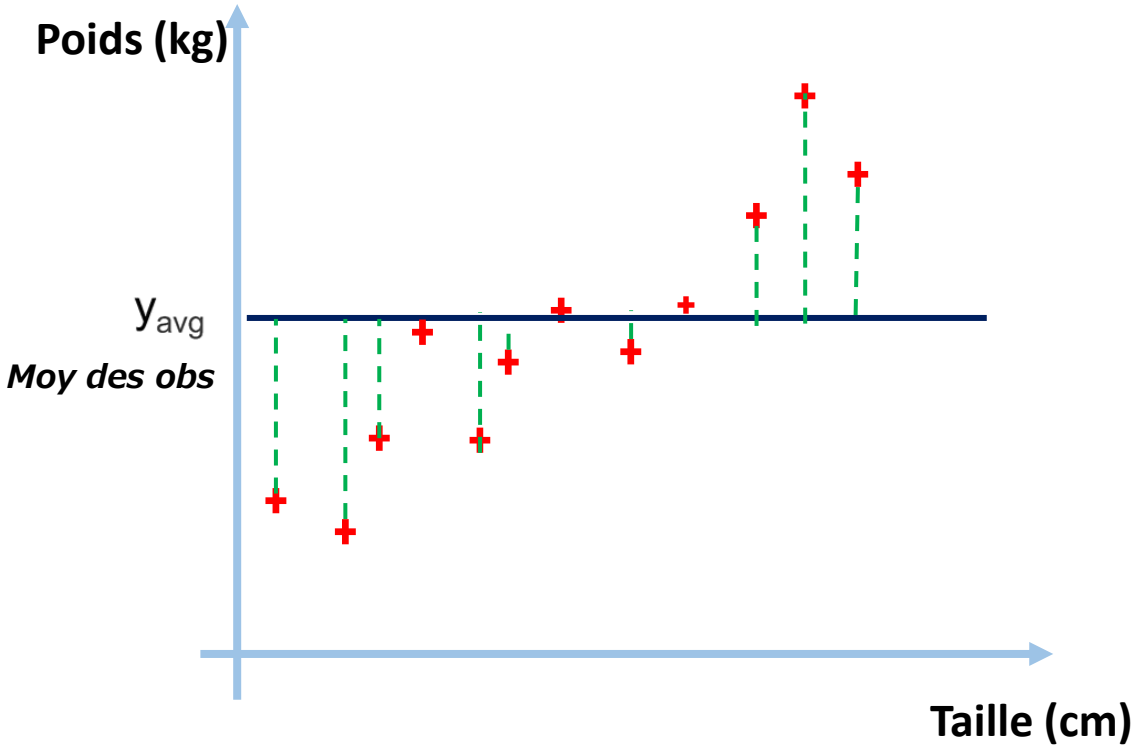


$$SS_{res} = \sum(y_i - \hat{y}_i)^2$$

Attention à la différence entre erreur et résidu !

Application à la régression linéaire simple

Le coefficient de détermination R²



$$SS_{tot} = \sum(y_i - y_{avg})^2$$

Erreur quadratique

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

variance

Il évalue la performance du modèle par rapport au niveau de variation présent dans les données

Application à la régression linéaire simple

méthode des moindres carrés ordinaires

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

Quelques remarques

- R^2 détermine la qualité de ma prédiction mais n'est pas le seul ! (critère d'Akaïke, critère de Schwartz, critère de Hannan Quinn...)
- **Valeur comprise en 0 et 1**
- Plus R^2 → 1 plus la droite de régression linéaire se rapproche de la droite d'observation
- R^2 est aussi appelé coefficient de détermination de pearson
- En python : .score

Application à la régression linéaire simple

Coefficient de détermination ou comment mesurer la qualité d'un ajustement

Problème dès que l'on rajoute des variables indépendantes :

R^2 va toujours augmenter... Même si l'ajout des variables n'est pas pertinente

R^2 sera biaisé!

$$\boldsymbol{y = b_0 + b_1 x_1 + \varepsilon}$$

$$\boldsymbol{y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + \varepsilon}$$

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

$$Adjusted\ R^2 = 1 - (1 - R^2)\ \frac{(n-1)}{(n-p-1)}$$

Avec n la taille de l'échantillon (nbr de points observés)
et p le nombre de variables indépendantes

Application à la régression linéaire simple

Autres métriques

Mean Square Error

→ Pénalise beaucoup plus les grandes erreurs que la MAE
→ Si on accorde une grande importance aux grandes erreurs

$$\text{MSE} = \frac{1}{2m}\sum(erreur)^2$$

Mean Absolute Error

→ Si le dataset contient des outliers (valeurs abbérantes)
→ Peu sensible aux grandes erreurs

$$\text{MAE} = \frac{1}{2m}\sum|(erreur)|$$

Pour aller plus loin :

- https://www.youtube.com/watch?v=_TE9fDgtOaE
- https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e
- https://towardsdatascience.com/which-evaluation-metric-should-you-use-in-machine-learning-regression-problems-20cdaef258e

Faisons le point

Nous avons appris :

- **La différence entre statistique descriptive et inférentielle**
- **Ce qu’est un estimateur**
- **Quelles sont les qualités d’un bon estimateur**
- **Ce qu’est la p-value**
- **Comment fonctionne la méthode des moindres carrés**
- **Les différentes métriques associées à la régression**

Application à la régression linéaire multiple

Rappel d’équation
Construction d’un modèle
Les différentes approches
application

Application à la régression linéaire multiple

Rappel de l'équation

$$y = a_0 + a_1 * x_1 + a_2 * x_2 + \dots a_n * x_n + \varepsilon$$

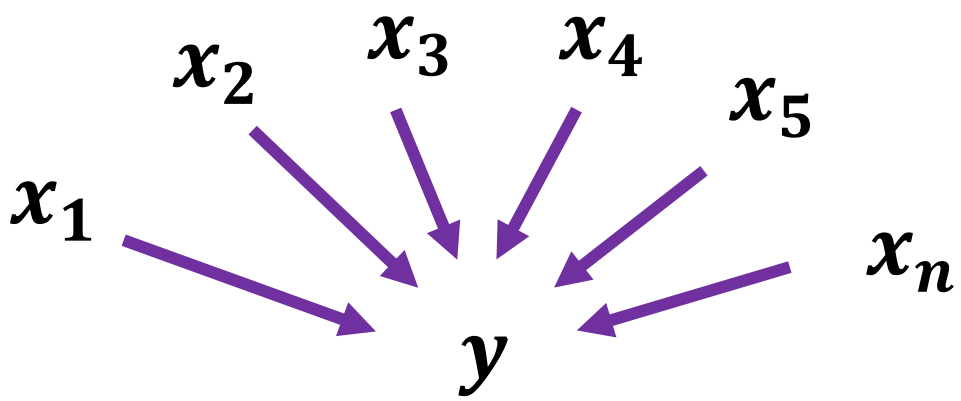
Sous forme matricielle :

$$\begin{pmatrix} y_1 \\ \\ y_i \\ \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & & x_{1p} \\ 1 & & & \\ 1 & x_{i1} & x_{ij} & x_{ip} \\ 1 & & & \\ 1 & x_{n1} & & x_{np} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \\ \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \\ \varepsilon_i \\ \\ \varepsilon_n \end{pmatrix}$$

$$Y = Xa + \varepsilon$$

Application à la régression linéaire multiple

Construction d'un modèle



Toutes les variables indépendantes peuvent être des prédicteurs potentiels de y

comment choisir ??

Pourquoi faut il forcément en sélectionner ??

Application à la régression linéaire multiple

Construction d'un modèle

5 méthodes :

- On construit un modèle en utilisant toutes les variables indépendantes
- Backward Elimination
- Fordward Elimination
- Bidirectionnal elimination

} Stepwise Regression

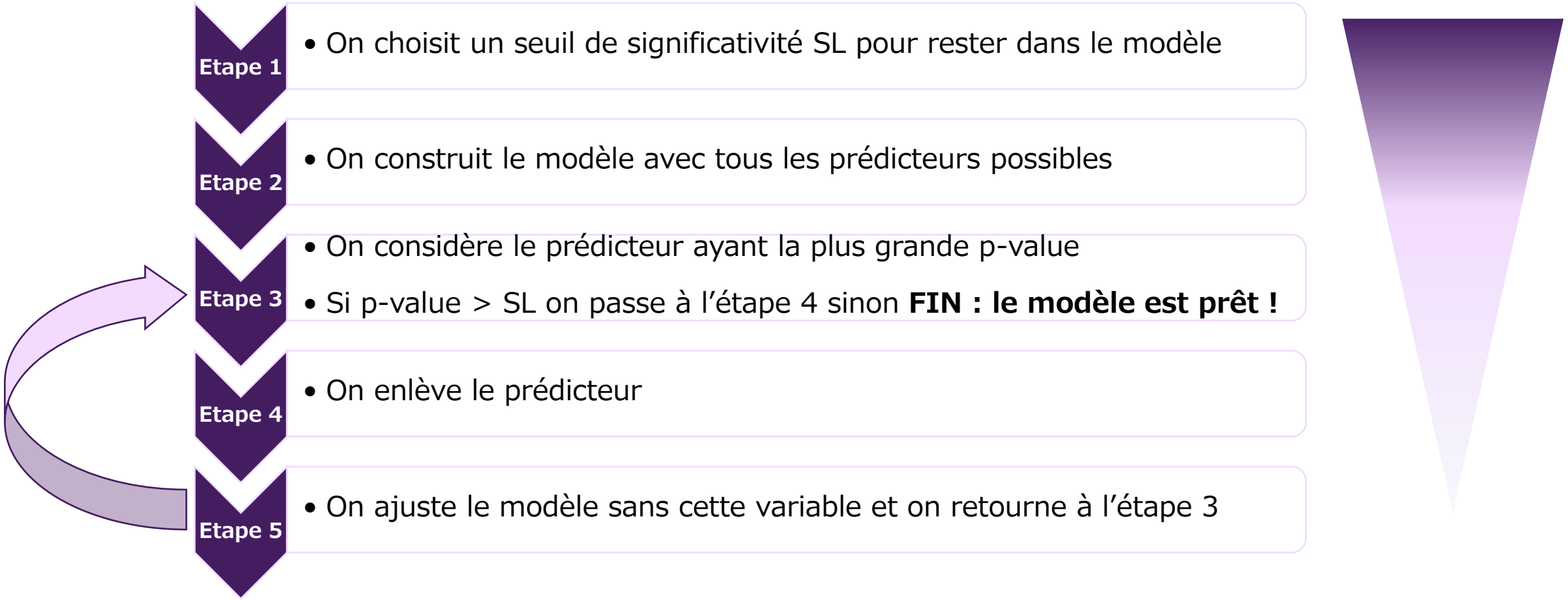
- On garde toutes les variables indépendantes et on construit tous les modèles possibles

Laquelle choisir ??

Application à la régression linéaire multiple

Les différentes approches

BACKWARD ELIMINATION



Application à la régression linéaire multiple

Les différentes approches

FORWARD ELIMINATION

Etape 1

- On choisit un seuil de significativité SL pour entrer dans le modèle

Etape 2

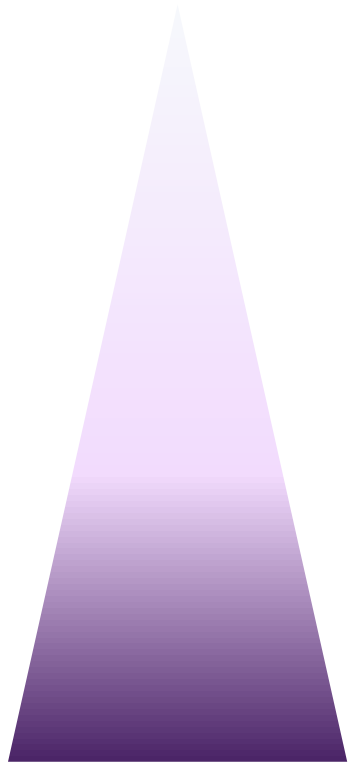
- On construit tous les modèles simples de régression $y \sim x_n$ et on sélectionne celui qui a la plus petite p-value

Etape 3

- On garde cette variable et on ajuste tous les modèles possibles avec un prédicteur en plus

Etape 4

- On considère le prédicteur ayant la plus petite p-value
- Si p-value<SL on retourne à l'étape 3 **sinon FIN : notre modèle est prêt ! Attention : on garde le modèle précédent !!!**



Application à la régression linéaire multiple

Les différentes approches

BIDIRECTIONNAL ELIMINATION

Etape 1

- On choisit un seuil de significativité SL pour entrer dans le modèle et un pour rester

Etape 2

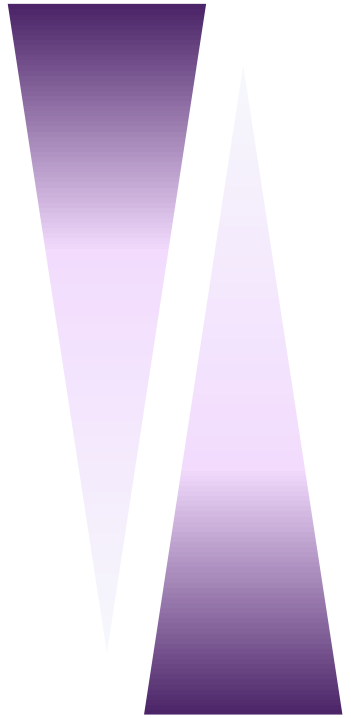
- On effectue l'étape suivante de la Forward selection à savoir :
- Les nouvelles variables doivent vérifier p< SLentrée pour entrer dans le modèle

Etape 3

- On effectue toutes les étapes de la Backward Elimination à savoir :
- Les anciennes variables doivent vérifier p< SLsortie pour rester dans le modèle

Etape 4

- Aucune nouvelle variable peut entrer ni aucune ancienne variable peut sortir **FIN : notre modèle est prêt**



Application à la régression linéaire multiple

Les différentes approches

On construit tous les modèles possibles

Etape 1

- On choisit un critère d'ajustement (ex : critère D'Akaïke)

Etape 2

- On construit tous les modèles de régressions possibles
- Soit 2^{n-1} combinaisons possibles

Etape 3

- On choisit celui dont le critère a la meilleure valeur possible

Pour aller plus loin

POUR aller plus loin :

- https://irma.math.unistra.fr/~fbertran/enseignement/Master1_2010_2/Master2Cours3.pdf
- https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce
- http://mathsv.univ-lyon1.fr/app/cours/?theme=proba&chap=7
- https://www.ibm.com/fr-fr/analytics/learn/linear-regression
- http://theses.univ-lyon2.fr/documents/lyon2/2004/legrand_g/pdfAmont/legrand_g_chapitre02.pdf
- https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

Quelques explications plus complètes sur la notion de p-value :

- https://www.youtube.com/watch?v=xVIt51ybvu0
- https://www.youtube.com/watch?v=jy9b1HSqtSk
- https://freakonometrics.hypotheses.org/2462

Faisons le point

Nous avons appris :

- **Les conditions préalables pour utiliser la régression linéaire multiple**
- **Les différentes approches pour la sélection des variables**

Application: Quelles seront les futures licornes de demain ?

Contexte du projet

La BPI France dispose d'un fond d'investissement qu'elle voudrait utiliser pour investir dans les start'up de demain les plus prometteuses. Seulement elle ne sait pas comment les sélectionner. Faut -il investir dans celles qui dépensent le plus en marketing ? en recherche et développement ? Dans quelles villes les startups semblent mieux opérer? Elle fait donc appel à vous pour y voir plus clair...

Modalités pédagogiques

Vous travaillerez en binôme sur ce projet. Vous aurez jusqu'au vendredi 18 au soir pour l'envoyer par mail à barbarafraenckel@gmail.com

Critères de performance

Concernant le notebook : En fournissant les budgets alloués aux différents pôles de travail, le modèle doit pouvoir prédire quel est le profit potentiel qui serait généré. Concernant le rapport écrit : une analyse construite , synthétique et méthodique. qui justifie l'utilisation d'une des approches de sélection de variables vues en séance et qui apporte des éléments de réponses aux questions de la BPI.

Livrables

Vous devez donc concevoir un modèle de régression linéaire multiple qui permettra à la BPI d'une part de sélectionner les 5 start'up les plus prometteuses et d'autre part de déterminer dans quel(s) secteur(s) il serait le plus judicieux de répartir les budgets de dépenses. Le livrable se présentera sous forme d'un notebook python ET d'un rapport écrit synthétisant votre démarche de travail que vous justifierez par les résultats des différentes analyses que vous aurez effectuées.

Les groupes

- Andréa & Audrey
- Fanny & Alice
- Anouar & Jauffrey
- Matis & Eric & Arthur
- Victor & Jordan N
- Jean & Lorraine
- Babacar & Soubika
- Thibaut & Aissa & Yanis