REGRESSION LINEAIRE

PART II

TABLE DES MATIERES

I. HYPOTHESES DE LA REGRESSION LINEAIRE

Avant d'appliquer un modèle de régression linéaire, il faut s'assurer que les cinq hypothèses suivantes soient vérifiées :

1. **Exogénéité** : cette hypothèse suppose que les variables explicatives ne sont pas corrélées en terme d'erreur. On dit que les variables explicatives sont exogènes.
2. **Homoscédasticité** : les termes d'erreurs sont supposés de variance constante.
3. **Erreurs indépendantes**
4. **Normalité des erreurs** : cette hypothèse suppose que les erreurs entre la valeur observée et la valeur prédite sont distribuées normalement. i.e : $y_i - \hat{y}_i \sim N(0,1)$
5. **Non colinéarité des variables indépendantes** : Cette hypothèse suppose qu'aucune des variables explicatives du modèle ne peut s'écrire comme une combinaison linéaire des autres variables.

Etant donné que le modèle de régression linéaire est généralement moins robuste que d'autres modèles de régression, nous ne nous attarderons pas sur ces notions.

Toutefois, si vous serez amené à élaborer de tel modèle, il faut absolument vérifier ces hypothèses dans premier temps en utilisant des tests d'hypothèses appropriés.

II. EVALUATION D'UN MODELE DE REGRESSION

Pour les problèmes de régression il existe plusieurs mesures pour évaluer la qualité d'un modèle. Elle se basent toutes sur des calculs réalisés à partir de deux grandeurs :

- La valeur observée d'une série à prédire y_i (valeur réelle)
- La valeur prédite par le modèle pour cette même valeur observée $\hat{y}_i$

Ces grandeurs permettent de définir des indicateurs de performance (ou métriques) du modèle.

Les plus connus sont les suivants :

1. MEAN ABSOLUTE ERROR (MAE)

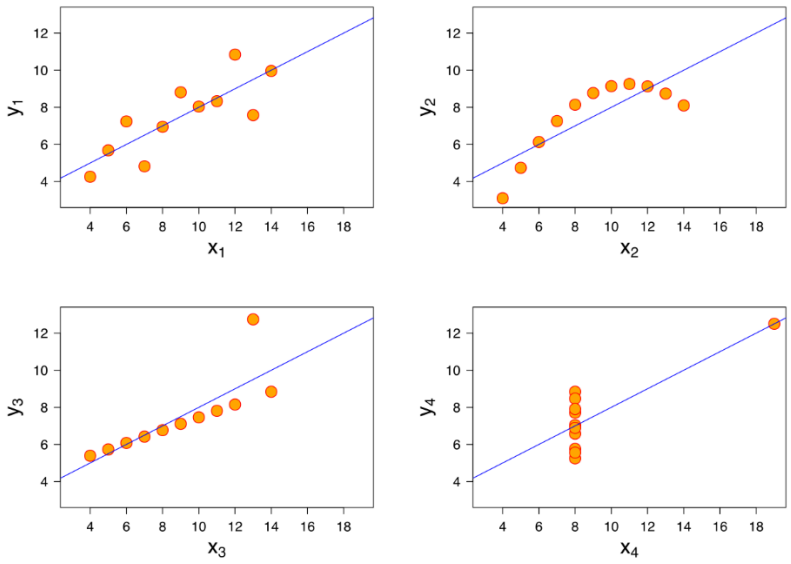
L'erreur moyenne absolue est exprimée sous la forme :

$$MAE = \frac{1}{m}\sum_{i=1}^{m} |y_i - \hat{y}_i|$$

La MAE correspond à une indication agrégée de l'erreur de prévision.

Le MAE ne pénalise pas les grandes erreurs d'où l'utilisation de l'erreur quadratique moyenne

Exemple : même modèle de régression linéaire avec la même valeur du MAE.



2. Mean Squared Error (MSE)

Il s'agit de l'erreur quadratique moyenne exprimée par :

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

Cependant, l'unité de la MSE n'est pas la même que l'unité de la variable expliquée y ce qui rend son interprétation difficile. D'où l'utilisation de la RMSE.

3. Root Mean Square Error (RMSE)

Il s'agit de la racine de l'erreur quadratique moyenne exprimée comme suit :

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

Exemple : pour modèle qui prédit la quantité de médicament à administrer, de petites fluctuations du RMSE peuvent être très importantes.

4. Coefficient de determination R^2

Le coefficient de détermination est donné par la formule suivante :

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

Le R^2 permet d'avoir une idée générale de la performance du modèle. En effet, il permet de comparer l'écart à la moyenne de la variable y à l'écart de la prévision.

Il peut donc être interprété comme la part de **la variation de y attribuable au modèle** ou comme **une mesure d'adéquation entre le modèle et les données observées.**

En d'autres termes, il mesure à quel point la droite de la régression linéaire se rapproche le mieux de la droite horizontale de la valeur moyenne.

Sa valeur est comprise entre 0 et 1 :

- 0 indique une adéquation nulle
- 1 indique une adéquation parfaite

Remarques :

- dans le cadre d'une régression linéaire simple, $R^2 = r^2$ (le carré du coefficient de corrélation de Pearson)

- R^2 peut être négatif si la somme des erreurs quadratiques est supérieure à la variance. Ce qui signifie que la droite de régression linéaire se rapproche moins bien des valeurs observées que la moyenne de ces valeurs.

Le problème avec l'utilisation de R^2 comme indicateur de performance pour évaluer deux modèles de régression linéaire n'est toujours pas pertinent surtout dans le cas de régression multivariée. En effet, le R^2 ne va jamais diminuer en ajoutant des variables explicatives puisque la variance reste inchangée. Ce qui va biaiser l'interprétation de la qualité du modèle d'où le recours au R^2 ajusté.

5. Adjusted R^2

$$R^2_{adjusted} = 1 - (1 - R^2)(\frac{m-1}{m-p-1})$$

Avec :

- m : taille de l'échantillon
- p : nombre de variables explicatives

On ajoute alors un facteur de pénalisation qui permet de prendre en considération l'augmentation de R^2 lorsqu'on ajoute une variable explicative au modèle de régression et la contrebalancer.

Le adjusted R^2 est donc une métrique très utilisée pour comparer les modèles de régression linéaire parce qu'elle permet de mesurer effectivement l'impact d'une variable explicative sur la variable expliquée.

6. Conclusions

- Pour interpréter toutes ses mesures, on les compare souvent à la moyenne de la variable expliquée pour avoir une idée sur la performance globale.
- Le MAE et RMSE ont l'avantage d'être dans l'unité de la variable à expliquer, ce qui permet d'en avoir une interprétation fonctionnelle.
- Par rapport au MAE, le MSE permet de punir plus sévèrement les grandes erreurs.
- Une technique souvent utilisée consiste à comparer la métrique d'erreur avec la moyenne
- Le contexte est très important pour le choix des métriques pour évaluer le modèle.
- La connaissance métier joue un rôle très important pour l'interprétation des indicateurs de performances.

III. Pratiquer

1. Dataset

50_startups.csv

L'obejctif est de prédire le profit des start up en se basant sur les dépenses dans différents domaines : R&D, admin et marketing, et l'état de la start up.

Autrement dit, on cherche à savoir quelles variables ont le plus d'impact sur le profit.

2. DATA PROCESSING

1. Transformer les variables catégorielles en variables numériques
2. Préparer le dataset de training et de test

3. MODELE DE REGRESSION LINEAIRE

1. Variables explicatives : `[0, 1, 2, 3, 4, 5]`
2. Variables explicative : `[0, 1, 3, 4, 5]`
3. Variables explicative : `[0, 3, 5]`
4. Variables explicative : `[0, 3]`

4. ÉVALUATION DES MODELES

En se basant sur les indicateurs de performances, évaluer les modèles construits précédemment.

IV. FEATURE SELECTION

La sélection des variables explicatives, ou autrement dit feature selection, est une étape très importante dans la construction d'un modèle de machine learning.

La feature selection nous permet de sélectionner les variables explicatives qui nous permettent d'avoir un 'bon' modèle :

- Si on supprime beaucoup de variables, on ne peut pas avoir un bon modèle. C'est ce qui est connu sous le nom de « **Garbage In, Garbage Out** ».
- Si on garde toutes les variables, il sera peu pratique, voire difficile, d'expliquer votre modèle à votre client et en quoi les variables explicatives utilisées ont impact sur la variable expliquée.

Les avantages des méthodes de sélection de variables sont multiples :

- Apprentissage plus rapide
- Diminution de la complexité
- Performances améliorées
- Réduction du overfitting

La sélection de variables peut être effectuée de plusieurs manières, mais on distingue trois catégories :

- Filter Method
- Wrapper Method
- Embedded Method

1. FILTER METHOD

Comme son nom l'indique, cette méthode filtre les variables explicatives pour garder que les variables pertinentes.

Ce filtre est basé sur une matrice de corrélation, le plus souvent la corrélation de Pearson pour garder les variables explicatives qui dépasse un certain seuil de corrélation avec la variable expliquée.

2. WRAPPER METHOD

Cette méthode consiste à ajouter et/ou supprimer des features en fonction des performances de l'algorithme.

Il s'agit d'une méthode itérative et couteuse en calcul mais plus précise que la méthode précédente.

Il existe plusieurs méthodes de type wrapper telles que :

- Backward Elimination
- Froward Selection
- Bidirectional Elimination
- Recursive Feature Elimination

Dans ce qui suit, seules les méthodes Backward Elimination et RFE.

BACKWARD ELIMINATION

Comme son nom l'indique, on considère le premier modèle avec toutes les features. Les features les moins performantes seront éliminées une par une jusqu'à ce que les performances globales du modèle se situent dans une zone d'acceptation.

La mesure utilisée pour évaluer la performance des features est la p_value.

Si on considère un seuil de 0.05, si la p_value est supérieur à ce seuil, la variable explicative est éliminée sinon elle est conservée.

Il s'agit d'un processus itératif.

RECURSIVE FEATURE ELIMINATION

La méthode RFE fonctionne en supprimant récursivement les feature. Elle utilise la métrique de précision pour classer les features en fonction de leur importance. Dans ce classement, une feature avec une valeur 1 est la variable explicative la plus importante.

sklearn.feature_selection.RFE

3. EMBEDDED METHOD

Les embedded méthodes permet de sélectionner à chaque itération la variable explicative qui contient le maximum d'information à cette itération. Les méthodes de régularisation sont les plus utilisées qui pénalise une feature selon un seuil.

Parmi ces méthodes, la méthode de régularisation Lasso.

Si la feature n'est pas pertinente, la méthode Lasso la pénalise en mettant à 0 son coefficient. Par conséquence, la feature est éliminée et le reste est conservée.

sklearn.linear_model.Lasso

sklearn.linear_model.LassoCV

Feature selection using SelectFromModel and LassoCV

4. CONCLUSIONS

- La Filter method est moins précises. Elle permet cependant de vérifier la multilinéarité des variables explicatives.
- Les métho

Il est évident qu'il existe bien d'autres méthodes de sélection de features, qu'on abordera au fur et à mesure de la progression de la formation.

V. PRATIQUER

1. DATASET

Boston

2. FEATURES SELECTION

1. Filter Method
2. Backward Elimination
3. Recursive Feature Elimination

3. EVALUATION DES MODELES

En se basant sur mes indicateurs de performances de la régression, évaluer les différences modèles élaborés.

4. INTERPRETATION

Interpréter la droite de régression obtenue avec le meilleur modèle.

N.B : Il est fortement conseillé de créer une libraire ou une classe avec les différentes méthodes de sélection des variables que vous pouvez réutiliser avec d'autres datasets.

VI. POUR LA CULTURE

Machine Learning — Introduction to Feature Selection and Backward Elimination

Intro to Linear Model Selection and Regularization

How to Perform Lasso and Ridge Regression in Python

sklearn.feature_selection: Feature Selection