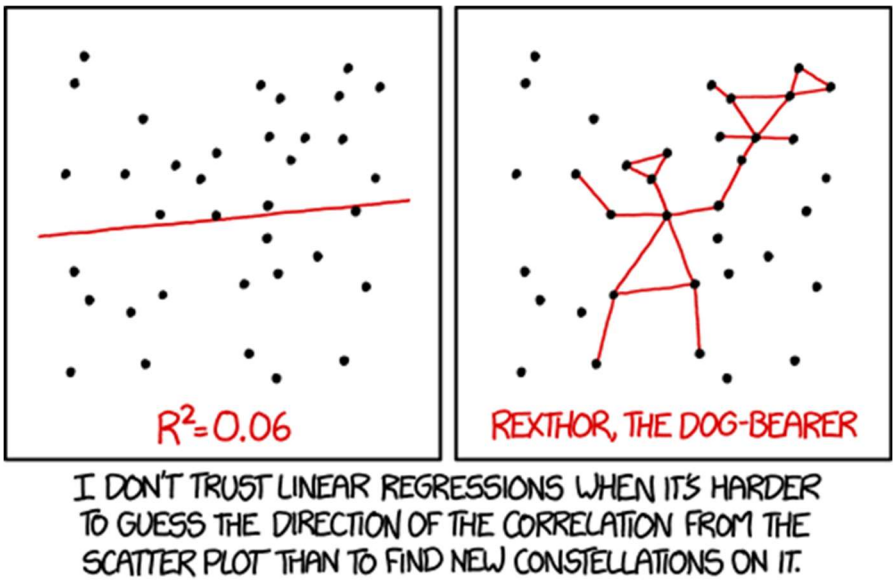
STATISTIQUES :

Régression linéaire



Linear Regression on xkcd

Source : https://xkcd.com/1725/

Objectifs pédagogiques

Appliquer les techniques de régression linéaire à l'aide de python et interpréter les résultats

La régression linéaire peut être utilisée à la fois pour comprendre la relation entre différentes mesures et pour utiliser un ensemble discret de mesures pour faire des prédictions sur une autre quantité. Dans ce module, nous examinerons les différences entre ces deux approches, et nous apprendrons comment appliquer chacune à l'aide des bibliothèques python appropriées. Nous le ferons en utilisant un certain nombre de jupyter notebooks qui contiennent des exercices à compléter. Ceux-ci construiront lentement votre intuition jusqu'à ce que vous soyez prêt à travailler sur vos propres projets de manière plus indépendante.

Un autre objectif de ce module est d'utiliser le temps pour développer votre compréhension du fonctionnement des méthodes statistiques. Cela vous permettra de les appliquer de manière plus appropriée et avec succès lors de la résolution de différents types de problèmes.

Compétences développées

- Savoir réaliser une régression linéaire avec Python
- Savoir interpréter les résultats d’une régression linéaire

Démarche pédagogique

- Durée du projet : 5 jours (petits projets de 1,5 + 1,5 + 2 jours).
- Ce projet sera réalisé par deux
- Partagez des conseils avec vos voisins !
- Tout au long des jupyter notebooks, il y aura du code manquant et des tâches qui doivent être terminées :
    - **Exercice :** Indique un travail essentiel qui vous permettra de valider les compétences
    - **Tâche :** Indique une section à compléter qui améliorera votre compréhension

Etape 1

Utiliser la régression linéaire pour estimer l'effet d'une variable (ou de plusieurs) sur une autre (1.5 jour)

Objectifs de l'activité

- Introduire la notion d'approches différentes de la modélisation statistique
- Introduire et utiliser les bibliothèques python pour faire une régression linéaire classique : scipy, statsmodels
- Apprendre à interpréter la sortie d'une régression linéaire, effectuer la vérification du modèle et la sélection des variables
- Utiliser la régression linéaire pour comprendre le surpeuplement des services d'urgence de l'hôpital, ainsi que les caractéristiques du vin (pas de connexion)

Compétences

- Sélectionner des approches appropriées pour la modélisation statistique
- Utiliser et interpréter la sortie d'une régression linéaire classique
- Implémenter la 'backward selection' pour sélectionner des variables ('features') dans un modèle

Consignes

- Terminez toutes les tâches et les exercices dans le jupyter notebook '01ClasicLinearRegression.ipynb'
- Prenez le temps de lire les sections de livres suggérées
- Référez-vous aux ressources si nécessaire (ci-dessous et partout dans les jupyter notebooks)

Ressources

- Livre en pdf : https://www.statlearning.com/
  Chapitre 1
  Section 2.0->2.1.3
  Section 3.0->3.2
  Section 3.3.3
  Section 3.4
- Statsmodels documentation :
  https://www.statsmodels.org/stable/index.html
  https://www.statsmodels.org/devel/regression.html
- Explication de la sortie récapitulative du statsmodels :
  https://www.youtube.com/watch?v=U7D1h5bbpcs
  https://www.graduatetutor.com/statistics-tutor/interpreting-regression-output/
- Explication des modèles vérification résiduelles:
  https://www.youtube.com/watch?v=eTZ4VUZHzxw

Livrables

- Exercices terminés dans le jupyter notebook

Pour aller plus loin

- Téléchargez cet article (et les données associées), voyez si vous pouvez reproduire l'étude ou voir ce que vous pouvez faire d'autre avec la régression linéaire https://bmjopen.bmj.com/content/8/5/e020296

Etape 2

Utiliser la régression linéaire pour la prédiction (1.5 jour)

Objectifs de l'activité

- Introduire et utiliser les bibliothèques python : scikit-learn
- Calculer et utiliser l'erreur telle que l'erreur quadratique moyenne (mean squared error)
- Comprendre l'utilisation des données d'entraînement ('training') et de 'test'
- Créer un modèle de prévision des prix des logements
- Introduire des termes clés dans la modélisation statistique

Compétences

- Calculer l'erreur d'un ensemble de prédictions
- Effectuer une régression linéaire multiple avec le package scikit-learn
- Décrire les termes clés de la modélisation statistique

Consignes

- Pour une introduction à sklearn, lisez et exécutez le code dans '02a-IntroductionScikitLearn.ipynb'
- Terminez toutes les tâches et les exercices dans le jupyter notebook '02b-LinearRegressionForPrediction.ipynb'
- Prenez le temps de lire les sections de livres suggérées
- Référez-vous aux ressources si nécessaire (ci-dessous et partout dans les jupyter notebooks)

Ressources

- Livre en pdf : https://www.statlearning.com/
  Section 2.2
  Section 2.1.4 -> 2.1.5
  Section 4.1 , 4.2
- Algorithmes d'apprentissage supervisé et non supervisé : https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
- Pourquoi nous utilisons les données d'entraînement et les données de test https://www.youtube.com/watch?v=EuBBz3bI-aA

Livrables

- Exercices terminés dans le jupyter notebook
- Un mémo qui contient une description de chacun des termes clés :

Terme	Description
intercept, slope,	

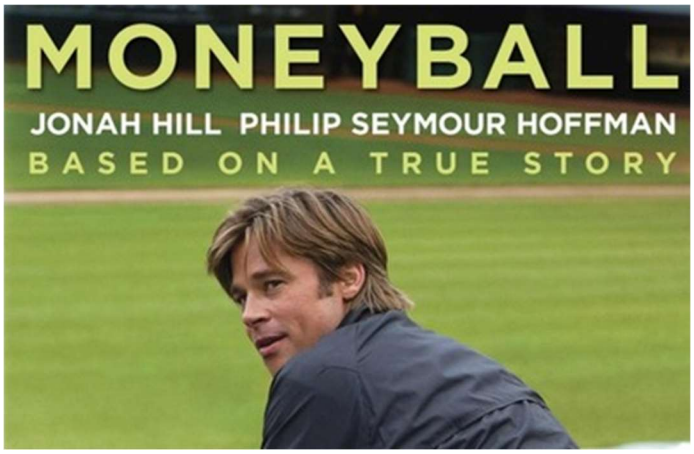
coefficients	
residual sum of squares	
supervised/unsupervised learning	
classification/regression	
parametric/ Non-parametric models	
overfitting/underfitting	
bias/variance tradeoff	
cross-validation	

Etape 3

Utiliser des techniques de validation croisée (*cross-validation*), de régularisation et de sélection de modèles (2 jours)

Objectifs de l'activité

- Introduire la validation croisée et comment elle peut être utilisée pour la sélection du modèle
- Introduire la notion de régularisation et les méthodes de régression linéaire qui l'utilisent
- Créer un modèle qui prédit les performances des joueurs de baseball



MONEYBALL- a true story of statistics in action
Source : https://www.3inno.com/

Compétences

- Utiliser la validation croisée k-fold
- Expliquer la validation croisée k-fold et la régularisation
- Utiliser un modèle de régression linéaire régularisé avec réglage d'hyperparamètre

Consignes

- Terminez toutes les tâches et les exercices dans le jupyter notebook '03a-CrossValidation&Regularisation.ipynb'
- Prenez le temps de lire les sections de livres suggérées
- Référez-vous aux ressources si nécessaire (ci-dessous et partout dans les jupyter notebooks)

Ressources

- Livre en pdf : https://www.statlearning.com/
  Section 5.1
  Section 6.0 (introduction only)
  Section 6.2

- Discussion sur les ‘Bias’ et ‘Variance’
  https://elitedatascience.com/bias-variance-tradeoff
- Summary of supervised learning process using scikit-learn
  http://ethen8181.github.io/machine-learning/model_selection/model_selection.html

Livrables

- Exercices terminés dans le jupyter notebook
- Produire un mémo qui explique (en vos propres termes) :
  - La procédure de validation croisée k-fold (inclure un schéma)
  - Les principes de régularisation
  - Les régressions de Lasso et Ridge