
TP : Régression linéaire

1. Consignes pour rendre votre travail : Le TP de chaque étudiant sera évalué par les pairs. Pour cela, vous devez déposer un unique notebook jupyter au format *.ipynb sur la rubrique dédiée (TP noté n°1) de la plateforme d'ecampus. Le nom du fichier doit être le vôtre, nom de famille uniquement (ex : Victor Dupont dépose le dossier `dupont.ipynb`). Le contenu du fichier doit être anonymisé : il ne doit comporter aucun signe susceptible de vous identifier.

Vous devez charger votre fichier, entre le lundi 11/10/2021 et le mardi 19/10/2021, 23h59.
Retard : malus de 4 pts par tranche de 24h (sauf excuses validées par l'administration).

2. Consignes pour la correction par les pairs : Entre le 20/10/2021 et le 27/10/2021 vous devrez corriger 2 copies qui vous seront assignées anonymement. **Ne divulguez à personne les numéros de copie qui vous ont été attribués.**

En pratique, vous devrez remplir 2 lignes d'une feuille de calcul (.xlsx) disponible sur le site pédagogique (document `autocorrection_TP1.xls` disponible sur ecampus), en indiquant le numéro de copie correspondant à chaque ligne. Pour chaque question vous devrez attribuer une note entre 0 et 2 :

- 0 (manquant/ non compris/ non fait/ insuffisant)
- 1 (passable/partiellement satisfaisant)
- 2 (bien)

Il faudra également évaluer de la même manière les points suivants (qui correspondent à 3 questions supplémentaires) :

- aspect global de présentation : qualité de rédaction, orthographe, présentation, graphes, titres, etc ...
- aspect global du code : indentation, lisibilité du code, commentaires adaptés
- Point particulier : absence de bug sur votre machine

Des commentaires pourront être ajoutés question par question pour aider la personne notée à s'améliorer, et de manière facultative mais obligatoire si vous ne mettez pas 2/2 à une question. Veuillez à rester polis et courtois dans vos retours.

Vous déposerez sur ecampus dans la rubrique dédiée un fichier .xlsx dans l'espace de rendu 'TP1 : correction par les pairs'. Les personnes qui n'auront pas rentré leurs notes avant la limite obtiendront zéro.

Rappel : aucun travail par mail accepté !

EXERCICE 1. (Analyse de la base de données "investment data") La lecture d'un tutoriel `pandas` pourra être utile : <http://pandas.pydata.org/pandas-docs/stable/tutorials.html>. Nous travaillons sur la base de données **Investment Data Set**¹ qui peut être téléchargée depuis https://bitbucket.org/portierf/shared_files/downloads/invest.txt. Avant de commencer, on réalisera l'exercice 12 du polycopié : "explicit formulas when $p = 1$ for prediction intervals", se trouvant dans le chapitre 3 : "Confidence intervals and hypothesis testing". On pourra aussi lire la section 3.3 de ce même chapitre du polycopié.

- 1) Importer la base de données "`invest.txt`" et l'afficher dans une forme lisible, *e.g.* une table contenant les 5 premières observations.
- 2) Réaliser le graphe suivant : la variable "Gross National Product" (GNP, column "`gnp`") est en abscisse et la variable "Investment" (column "`invest`") est en ordonnée. Transformer les 2 variables précédentes en échelle logarithmique. Nous travaillerons désormais avec les 2 nouvelles variables.

1. Voir Greene (2012) - *Econometric Analysis*, Prentice Hall, Upper Saddle River, NJ.

NOTE : Lorsque l'on traite des données monétaires, on travaille souvent en échelle logarithmique (pour prendre en compte les différences d'échelle).

Les questions suivantes (3 à 6) doivent être réalisées par l'intermédiaire d'opérations élémentaires, sans utiliser de bibliothèques existantes.

- 3) Pour la régression de "Investment" (variable à expliquer, output) sur "GNP" (variable explicative, covariable), estimer l'intercept et la pente, leurs écart-types, ainsi que le coefficient de détermination. Les afficher dans une forme lisible.
- 4) La pente estimée précédemment est-elle statistiquement significative ? On fera un test de student (t -test). Donner la valeur de la statistique de test ainsi que la p -valeur.
- 5) Pour GNP égal à 1000, estimer l'investissement prédit par le modèle. Pour GNP égal à 1000, donner l'intervalle de confiance pour la valeur prédite et l'intervalle de confiance pour la variable à expliquer "Investment", au niveau 90%. On pourra se référer à la section 3.1.3 "Confidence intervals for the predicted values" du polycopié dans laquelle chaque intervalle est défini, $CI(z)$ et $PI(z)$, respectivement (avec les notations du polycopié, $z = (1, 1000)^T$).
- 6) Sur un graphe avec échelle logarithmique, avec GNP en abscisse et investment en ordonnée, tracer les données, la droite de régression, ainsi que les intervalles CI et PI (pour toutes les valeurs de $\log(\text{GNP})$ comprises entre le maximum et le minimum observé sur les données)
- 7) En utilisant des classes/bibliothèques existantes, donner l'intercept, la pente, le coefficient de détermination ainsi que l'investissement prédit par le modèle quand GNP vaut 1000. La classe `LinearRegression()` de `sklearn.linear_model` est suggérée mais pas obligatoire. Vérifier que les valeurs calculées ici coïncident avec celles des questions précédentes.
- 8) Sur un graphe avec échelle logarithmique, avec GNP en abscisse et investment en ordonnée, tracer les données, la droite de régression, ainsi que l'investissement prédit par le modèle quand GNP vaut 1000 (on donnera à ce point une couleur différente).

NOTE : On introduit une nouvelle variable explicative, la variable **interest** (sans transformation logarithmique). Les questions suivantes (9 à 12) doivent être réalisées par l'intermédiaire d'opérations élémentaires, sans utiliser de bibliothèques existantes (on utilisera par exemple `inv` et `eig` de `numpy.linalg`).

- 9) Pour la régression de **Investment** sur GNP et **interest**, calculer la matrice de Gram non standardisée $Z^T Z$. Est-elle de rang plein ?
- 10) Pour la régression de **Investment** sur GNP et **interest**, estimer les 3 coefficients et leurs écart-types ainsi que le coefficient de détermination. En plus, faire un test de Student de significativité de chaque coefficient (donner la statistique de test et la p -valeur). Afficher les résultats dans une forme convenable. Discuter de la significativité des coefficients.
- 11) Pour les valeurs de GNP 1000 et **interest** 10, i.e., $z = (1, 1000, 10)^T$, prédire $\log(\text{investment})$ et donner les intervalles de confiance $CI(z)$ et $PI(z)$ au niveau 99.9%.
- 12) Sur un même graphe à 3 dimensions avec les axes suivants : $\log(\text{GNP})$, **Interest**, and $\log(\text{Investment})$, tracer les données, le "plan" de régression et les surfaces correspondantes aux intervalles de confiance à 99.9% (ces surfaces seront tracées sur le domaine de définition des données). On pourra par exemple utiliser la bibliothèque `mplot3d`
- 13) En utilisant des classes/bibliothèques existantes, donner les coefficients de régression, le coefficient de détermination ainsi que l'investissement prédit par le modèle quand GNP vaut 1000 et **interest** 10. La classe `LinearRegression()` de `sklearn.linear_model` est suggérée mais pas obligatoire. Vérifier que les valeurs calculées ici coïncident avec celles des questions précédentes.