Telecom Paris **MDI720**
2019/2020 **François Portier, Anne Sabourin**

TP1

On utilisera `IPython Notebook` pour réaliser ce TP. On déposera son compte-rendu sur le site pédagogique avant le 6 octobre, 23h59.

Le mot "régression" a été introduit par Sir Francis Galton (cousin de C. Darwin) alors qu'il étudiait la taille des individus au sein d'une descendance. Il tentait de comprendre pourquoi les grands individus d'une population semblaient avoir des enfants d'une taille plus petite, plus proche de la taille moyenne de la population ; d'où l'introduction du terme "régression". Dans la suite on va s'intéresser aux données récoltées par Galton.

1. Récupérer les données du fichier `https://bitbucket.org/portierf/shared_files/downloads/Galton.txt` (voir aussi leur description ici `http://www.randomservices.org/random/data/Galton.html`) et charger les avec `Pandas`.
2. Créer une colonne supplémentaire appelée "MeanParents" qui contient la taille du parent "moyen", c'est-à-dire $\frac{1}{2}$ (taille(pere) + 1.08taille(mere)). Pour plus d'explication sur cette transformation, on pourra consulter :
   https ://17art.ru/fr/calculator-of-the-ratio-of-height -and-weight-of-the-child-calculate-the-final-growth-of-the-child/

   On note x_i la taille du parent moyen pour la famille i et y_i la taille de l'enfant. On écrit $y_i = \theta_1 x_i + \theta_0 + \varepsilon_i$ et on modélise les variables ε_i comme centrées, indépendantes de même variance σ^2 inconnue.

3. Tracer le nuage de points (x_i, y_i) pour $1 \leq i \leq n$ où n est le nombre d'observations figurant dans les données.
4. Estimer θ_0, θ_1, par $\hat{\theta}_0$, $\hat{\theta}_1$ en utilisant la fonction `LinearRegression` de `sklearn`. Calculer et visualiser les valeurs prédites $\hat{y}_i = \hat{\theta}_1 x_i + \hat{\theta}_0$ et y_i sur un même graphique.
5. Vérifier la formule vue en cours liée au recentrage des données. On centrera les données et on vérifiera que la prédiction effectuée dans le modèle centré est bien la même que celle effectuée précédemment. Justifiez votre réponse. On pourra définir $y_{i,c}$ et $x_{i,c}$ comme étant les variables centrées et calculer

$$\hat{\theta}_c \in \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (y_{i,c} - x_{i,c}\theta)^2.$$

6. Visualiser l'histogramme des résidus $r_i = y_i - \hat{y}_i$ ($\hat{y}_i$ est la valeur prédite par le modèle). L'hypothèse de normalité est-elle crédible ?
7. Sur un graphique similaire à celui de la question 3, sur lequel apparait le nuage de points et la droite de régression, distinguer (à l'aide de deux couleurs différentes) les filles des garçons. Expliquer la forme de la distribution des résidus observée à la question précédente.
8. Effectuer une régression sur le groupe des garçons uniquement. Représenter cette droite de régression sur le graphique de la question précédente. Représenter la distribution des résidus obtenue sur ce modèle. Commenter.

9. A l'aide d'un partitionnement aléatoire de l'échantillon, construire un échantillon de test et un échantillon d'apprentissage. L'échantillon test doit contenir 20% des données. On estimera, sur l'échantillon d'apprentissage deux modèles : un premier modèle basé uniquement sur les garçons, et un deuxième modèle basé sur l'échantillon complet. On comparera les deux prédicteurs sur le groupe "garçon" de l'échantillon test en calculant le risque de prédiction :

$$\frac{1}{|\text{test}|} \sum_{i \in \text{test}} (y_i - \hat{y}_i)^2,$$

   $\hat{y}_i$ est la valeur prédite par le modèle.

10. En combinant les deux prédicteurs, celui appris sur les filles et celui appris sur les garçons, calculer le risque de prédiction associé. Pour chaque indicidu "garçon" on applique le modèle appris sur les garçon, pour chaque individu "fille" on lui applique le modèle appris sur les filles. On calculera ensuite le risque prédictif.
11. En utilisant, un encodage booléen de la variable `GENDER` construire le même prédicteur que précédement. Calculer, "à la main", son coefficient de determination. On comparera ce dernier au coefficient de determination du modèle avec seulement la variable `MeanParents`. On pourra se référer au polycopié pour la définition du coefficient de determination (dit aussi le R2).
12. Faire une régression avec les 2 variables explicatives suivantes : la taille du père et de la mère. Faire un graphique (en 3 dimensions) qui représente le nuage de point et le plan de régression. Visualiser la distribution des résidus.
13. A l'aide d'un échantillon test (construit comme précédemment), comparer la performance du modèle précédent à celle du modèle contenant seulement la variable `MeanParents`. Commenter.