Année 2020-2021 **M2 Data Science**

Statistique asymptotique
TD/TP 1 : Régression de Poisson, pénalisation et sur-dispersion

Avant de commencer,

- récupérer sur ma page le fichier `R_TP2.Rmd` et les données "bike_sharing".
- **conservez bien vos codes car nous allons les réutiliser pendant le TP3**

Vous trouverez une description des données sur https ://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset. Les données contiennent 731 heures d'observations (sur la période 2011-2012) du système de partage de vélo "Capital Bikeshare system", à Washington D.C. (USA). Pour chaque jour sont enregistrés les variables suivantes :

- `instant`: l'identifiant
- `dteday` : la date
- `season` : la saison (1 : printemps, 2 : été, 3 : automne, 4 : hiver)
- `yr` : l'année (0 : 2011, 1 :2012)
- `mnth` : le mois (1 to 12)
- `holiday` : 1 si c'est un jour férié (http ://dchr.dc.gov/page/holiday-schedule)
- `weekday` : le jour de la semaine (0 : dimanche, 1 : lundi, etc)
- `workingday` : 1 si le jour n'est ni un weekend, ni un jour férié, 0 sinon
- `weathersit` :
  - 1 : Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2 : Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3 : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- `temp` : la température en celsius normalisée $((t - t_{\min})/(t_{\max} - t_{\min})$, avec $t_{\min} = -8$ et $t_{\max} = +39)$
- `atemp` : a température ressentie en celsius normalisée $((t - t_{\min})/(t_{\max} - t_{\min})$, avec $t_{\min} = -16$ et $t_{\max} = +50)$
- `hum` : l'humidité normalisée
- `windspeed` : la vitesse du vent normalisée
- `casual` : le nombre de locations par des utilisateurs occasionnels
- `registered` : le nombre de locations par des abonnés
- `cnt` : le nombre de locations (incluant les utilisateurs occasionnels et les abonnés)

On veut expliquer et prédire la variable `cnt`.

Exercice 1

1. Charger les données et vérifier le nombre d'observations et de variables ainsi que leur type. Faire les changements de type nécessaires. Vérifier le lien entre les variables `casual`, `registered` et `cnt`. Que doit-on faire ?

2. Créer un jeu de données d'apprentissage `bike_sharing_train` contenant les informations pour des 585 premiers jours et un jeu de données de test `bike_sharing_train` contenant les informations des jours 586 à 731.

Exercice 2

Pour construire notre modèle, on travaille dans cette partie sur le jeu de données d'apprentissage seulement.

1. Représenter graphiquement la distribution de la variable `cnt`. Quelle modélisation proposez-vous ?
2. Etudier graphiquement le lien entre `cnt` et les variables explicatives. Quelles variables excluez-vous et pourquoi ?
3. Certaines variables explicatives sont liées. Lesquelles ?
4. Suite aux observations précédentes, proposez un premier modèle.

Exercice 3

1. Faire un premier modèle poissonnien `model_full`, en prenant en compte toutes les variables explicatives potentielles.
2. Calculer les prédictions obtenues avec le modèle sur les individus du train et du test, puis calculer la MAPE et le log-MSE (il faudra coder une fonction qui renvoie ces mesures sur le jeu de données de train et de test).
$$\log - \text{MSE} = \frac{1}{n} \sum_i (\log(\hat{Y}_i^P) - \log(Y_i))^2$$
3. Faire le test de nullité simultanée des coefficients des variables explicatives.
4. Faites un choix de modèle par AIC backward, on appellera le modèle obtenu `model_aic`.
5. Calculer la MAPE et le log-MSE du modèle. A l'avenir, conserver dans un même tableau les valeurs de ces mesures (sur le train et sur le test) dans un même tableau.

Exercice 4

1. Recommencer l'analyse en considérant les interactions d'ordre 2. Faites une sélection de variable par AIC backward puis comparer le modèle obtenu `model_inter` à celui de la question précédente. Attention à bien sauvegarder ce modèle grâce à `save(fit_aic,file = "./fit_aic")` puis `load("./fit_aic")` pour les recharger ensuite. On appellera le modèle obtenu `model_inter_aic`.

   Attention à bien vérifier que la `stepAIC` n'a supprimé un effet principal qu'à la condition qu'il n'intervienne plus dans des interactions.

Exercice 5

1. Calculer les MAPE et log-MSE des estimateurs ridge `model_ridge`, Lasso `model_lasso`, elastic-net `model_enet` (cross-validés via caret)
2. Pour le lasso, refaire un modèle non-pénalisé en ne considérant que les colonnes dont les coefficients Lasso sont non-nuls, on l'appelera `model_refit_lasso`
3. Dans le modèle refit du Lasso, faire une recherche d'individus aberrants et isolés. Enlevez vous des individus de l'étude ?

Exercice 6

Dans le modèle refit du lasso, changer la famille pour une quasi-poisson. Quel est l'estimation du paramètre de sur-dispersion ? Quelle conséquence cela a les intervalles de confiance et donc sur la sélection de variable ? Refaire une recherche d'individus aberrants et isolés, la conclusion change-t-elle ?