MST2 : Projet de Statistique

Nicolas Brunel

April 2020

Objectif Le but du projet statistique et de vous permettre de mettre en oeuvre les concepts du cours, à l'aide du langage R. Le travail se fait par binôme, et devra être rendu pour le 25 mai.

Travail à rendre

1. Un rapport rédigé de 10 pages maximum (avec des figures de taille raisonnable). La qualité de la rédaction, la pertinence et la justesse des analyses statistiques, et des conclusions seront centrales dans l'appréciation de ce travail.
2. Des codes R, commentés, en markdown (.rmd).
3. 5 slides de présentation et synthèses des résultats.

Le sujet Nous allons analyser les données "Rain in Australia" accessible sur le site Kaggle : https://www.kaggle.com/jsphyg/weather-dataset-rattle-package. Les données y sont décrites de manière détaillée et contiennent des informations météorologiques pendant 10 ans, dans plusieurs villes d'Australia. Les données sont journalières, et sont par exemple, les températures minimales ou maximales de la journée, la pression, l'ensoleillement,...

L'objectif de ce jeu de données est de prédire si il pleuvra le lendemain, à partir des variables disponibles le jour même. Il faut prédire RainTomorrow (Vrai ou Faux), mais attention il ne faut pas utiliser la variables Risk-MM, qui est une variable calculée, à partir du futur, donc inaccessible en pratique.

Phase 1 : Prise en main des données

1. Télécharger les données du site kaggle, et les charger dans un dataframe sous R Studio.
2. Faire une analyse descriptive des données : combien de variables différentes, leur type (quantitatif, qualitatif). On fera attention aux données manquantes (NA). On proposera une ou plusieurs stratégies de traitement de

ces variables manquantes : élimination ou imputation des valeurs manquantes (remplacement par la moyenne, médiane calculée ville par ville, en tenant compte du mois?...)

3. Faire une analyse descriptive : nombre de modalités, fréquence, moyenne, dispersion, représentation des distributions (boxplot, histogramme, tableau de fréquence,....)

4. Faire une analyse bivariée : calculer la covariance et la corrélation entre les variables continues. Tracer un diagramme de dispersion "scatter plot" pour les visualiser.

5. Théoriquement, quelle est la covariance entre une variable continue réelle et une variable discrète (par exemple prenant la valeur 0 ou 1). En pratique, comment la calculez vous et la représentez vous ? Représentez la pour un couple continue - discret.

6. Quelle est l'équivalent de la covariance entre deux variables catégorielles / qualitatives ?

Nous n'allons pas exploiter les variables possédant un trop grand nombre de valeurs manquantes : Evaporation, Sunshine, Cloud9am, Cloud3pm. Les études que nous ferons par la suite se feront ville par ville. Vous sélectionnez 3 villes pour lesquelles, nous n'avons pas plus de 5% de valeurs manquantes (à compléter) dans chaque variable.

Phase 2 : Modélisation des lois

Les données que nous avons sont des séries temporelles, et donc les observations ont donc une forte corrélation entre elles. Par exemple, si la variable T_n^{Max} représente la température maximale le jour n (dans la ville de Perth par exemple), nous ne pouvons pas dire que les données $(T_1^{Max}, \ldots, T_n^{Max}, \ldots, T_N^{Max})$ forment un échantillon i.i.d : nous avons une corrélation non-nulle entre les 2 variables aléatoires T_n^{Max}, T_{n+1}^{Max}. Et bien sûr en fonction de la saison, la distribution des températures évoluent et la densité n'est plus forcément la même.

Afin de se rapprocher un petit peu du cas i.i.d., nous allons faire une sous-échantillonnage pour avoir des données dont la loi est à peu prés similaire et indépendante. Nous allons modéliser la distribution des variables mois par mois, pour les 3 villes que vous avez sélectionné. Notre problématique statistique sera par exemple : "trouver la loi de la température maximale journalière à Perth, au mois de janvier". On supposera que les données sur 10 ans sont stationnaires...pas de polémique ici sur le réchauffement climatique. De telles études sur l'évolution sur 10 ans sont hors de portée de ce cours, et surtout nous supposons que ces variations sont d'un ordre de grandeur plus faible que les variations mensuelles et géographiques.

1. Afin de limiter la dépendance entre les données de jours successifs : Sélectionner aléatoirement 5 jours parmi chaque mois, sur les 10 années à votre disposition. Vous vous assurerez de ne sélectionner que des observations distantes d'au moins 3 jours. Vous obtenez ainsi 12 échantillons de taille 50.

2. Tracer l'histogramme, histogramme lissée, et proposer une distribution paramétrique pour les différentes variables (normal, gamma,...) continue. Faire un test d'adéquation, et sélectionner le meilleur modèle.

3. Ecrire la vraisemblance dans le cas paramétrique pour chacune des variables, et donner l'estimateur du maximum de vraisemblance des paramètres. Donner aussi les intervalles de confiance pour les paramètres.

4. Pour les différentes variables, tracer l'évolution des paramètres estimées en fonction du mois, dans les 3 différentes villes. Voyez vous des tendances ?

5. Pour confirmer ces tendances (ou au contraire confirmer la stationnarité), proposer des tests statistiques, et donner la statistique du test associé en vous basant sur le modèle paramètrique sélectionné.

6. Pour comparer les mois, nous allons considérer que nous faisons un test uniquement sur la moyenne de distribution du mois. Ecrire le test de comparaison de l'échantillon.

Phase 3 : Prédiction de la pluie

Nous allons faire un modèle prédictif pour les 3 villes sélectionnées.

1. Donner la probabilité mois par mois de pluie le lendemain. Ecrire le modèle statistique associée, et donner l'estimateur du maximum de vraisemblance associée et donner l'intervalle de confiance.

2. Faire un test statistique pour savoir si cette probabilité est différente de 5%, pour chaque mois.

3. Faire un test statistique pour comparer cette probabilité entre un mois d'hiver et un mois d'été. Peut on dire que la probabilité varie en fonction de l'année ? Entre deux villes?

4. Proposer un modèle prédictif basé sur une régression logistique : on modélise la probabilité conditionnelle de pluie le lendemain, sachant les variables que nous avons observées. Ecrire la vraisemblance associée.

5. Avec la fonction glm, estimer un modèle prédictif. Donner les valeurs des paramètres et les intervalles de confiance. A votre avis, comment sont ils estimés ?

6. Tester l'efficacité de votre prédicteur, et estimer votre erreur de prédiction à l'aide des jours que vous n'avez pas utilisés.

7. Y a t il une différence dans les modèles estimés dans les 3 villes ? Est ce qu'un modèle estimé dans une ville marche bien dans une autre ? Est ce que les paramètres ont le même signe, comment est ce que cela s'interprète ?

8. Combiner les données des 3 villes pour faire un nouveau modèle prédictif? Quels sont ses performances ?

Annexes

Régression logistique

Soit Y une variable à prédire et $X = (X_1, ..., X_n)$ des variables explicatives, le but de la régression est de trouver la relation qu'il y a entre X et Y c-a-d trouver f tel que $Y = f(X)$ (ex: regression linéaire, on pose $f(X) = \sum_i B_i X_i$ et on estime B).

La régression logistique est une extension du modèle linéaire pour des variables réponse Y de type catégoriel (Oui/Non). Y étant discret, l'estimer directement devient difficile. Ainsi, le problème est légèrement modifié, on ne cherche plus à prédire Y mais la probabilité $\mathbf{P}(Y|X)$. Un bon candidat de fonction est la logistique $f(x) = \frac{e^x}{1+e^x}$ (voir fig 1)

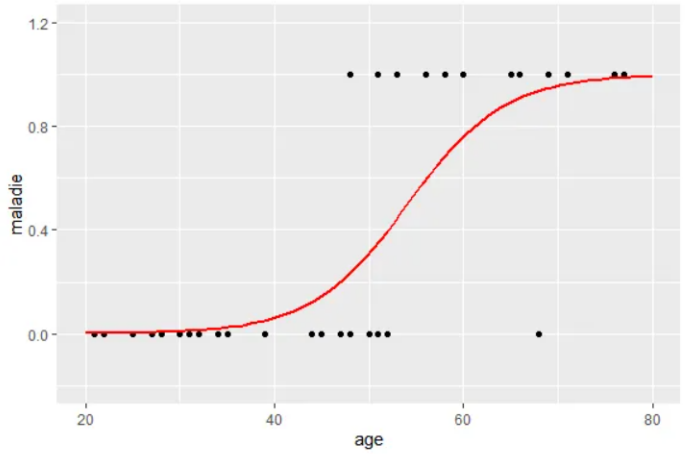

Figure 1: Estimation avec la fonction logistique

On peut alors ajuster les données à la fonction logistique:

$$f(X) = \mathbf{P}(Y = 1|X) = \frac{e^{\sum_i B_i X_i}}{1 + e^{\sum_i B_i X_i}} \quad (1)$$

Remarque: cette modélisation est équivalente à supposer que le rapport des probas ($\mathbf{P}(Y = i|X)$ i=1,0) suit un modèle linéaire:

$$\log(\frac{\mathbf{P}(Y = 1|X)}{\mathbf{P}(Y = 0|X)}) = \sum_i B_i X_i \tag{2}$$

Estimation des paramètres

Supposons que nous disposons d'une seule variable explicative et un échantillon $\{(X_1, Y_1), ..., (X_n, Y_n)\}$, d'après notre modélisation $Y_i \sim \mathcal{B}(f(X_i))$ avec $f(X_i) = \frac{e^{BX_i}}{1+e^{BX_i}}$. On peut alors estimer le paramètre B par maximum de vraisemblance:

$$B^* = \arg\max_B L(Y_1, .., Y_n|B) = \arg\max_B \prod_i f(X_i)^{Y_i}(1 - f(X_i))^{1-Y_i} \tag{3}$$

Sous R

Pour construire un modèle logistique, il faut utiliser la library **glm** qui se trouve dans le **package('stats')**. Pour estimer les paramètres, on appelle la fonction **glm.fit()** en lui donnant les données (dataFrame), la variable target (Y), les variables explicatives (X_i) et le modèle "binomiale".

```
> mylogit = glm(target ~ var1 + var2 + var3,
data = mydata, family = "binomial")
```

Pour la prédiction, on utilise la fonction **glm.predict()** en lui donnant les données (newData) et en précisant le type "response" pour avoir les probabilités.

```
> mylogit = glm.predict(newData, type="response")
```

Test d'hypothèse

Le principe d'un test d'hypothèse est de répondre de façon binaire à une question sur le paramétre en jeu. On se ramène alors à deux hypothèses $\mathbf{H}_0$ et $\mathbf{H}_1$.

En pratique, on se repose sur la p-value qu'on résume ainsi: "La p-value est la probabilité, sous $\mathbf{H}_0$, d'obtenir une statistique de test au moins aussi extrême que celle observée. La règle de décision est alors:

- Une petite p-value $\leq \alpha$ ($\alpha \approx 0.05$) nous dit que la statistique de test (données) est significativement invraisemblable sous $\mathbf{H_0}$, on rejette donc $\mathbf{H_0}$.
- Une p-value suffisamment grande nous dit juste qu'on ne peut pas rejetter $\mathbf{H_0}$. (**ça ne veut pas dire qu'elle est vraie, on ne sait pas...**)

Sous R

En pratique, on peut regrouper les tests en trois groupes. Les tests type student (**t.test** sur R) basé en général sur des statistiques de la forme $t = (m - \mu)/\sigma\sqrt{n}$ qui permettent de comparer des moyennes entre 2 échantillons.

- Hypothèse $\mathbf{H}_0$: $\mu_1 - \mu_2 = 0$ et $\mathbf{H}_1$: $\mu_1 - \mu_2 \neq 0$
- Pour l'utiliser, il faut lui donner les deux échantillons et les hypothèses (indépendance ? Variance egale ?) ex :

```
> data <- data.frame(var1, var2, var3);
> t.test(data$var1, data$var2, var.equal=FALSE, paired=FALSE);
Welch two-sample t-test
data: data$var1 and data$var2
t = -6.0315, df = 197.35, p-value = 7.88e-09
alternative hypothesis: true difference in means is not equal
to 0
```

Remarque: Il existe une généralisation de ce test, pour comparer plusieurs groupes en même temps. C'est le test ANOVA (**aov** sur R).

Le test chi-square (**chisq.test** sur R) pour comparer deux variables catégorielles.

- Hypothèse $\mathbf{H}_0$: Les deux variables sont indépendantes, ex:

```
> data.table <- table(data$var1, data$var2);
> data.table;

        chocolate strawberry
Female          1          4
Male            3          2
> chisq.test(data.table);
Pearson's chi-squared test with Yates' continuity correction
data: data.table
X-squared = 0.41667, df = 1, p-value = 0.5186
```

Si on suppose que les observations ne sont plus gaussiennes et qu'on dispose d'un grand échantillon, on peut utiliser les tests non-paramétriques. Par exemple Wilcoxon signed (**wilcox.test** sur R) pour comparer deux distributions. Noter qu'en général, les tests paramétriques (t-test, chi-square, anova) marche mieux.