Modele Lineaire

Problématique et présentation des données

Déterminer le risque qu'un sinistre se produisant en ayant à dispositions un jeu de données est au coeur de l'assurance non vie.

Nous étudierons ici le jeu de données "Pluviométrie dans les villes françaises" provenant de https://husson.github.io/data.html: c'est un tableau Excel recensant des données météorologiques sur 34 villes françaises en 2023.

Dans le contexte de notre base de données, étudier l'ampleur des précipitations a un intérêt pour par exemple fixer le prix d'un contrat d'assurance habitation.

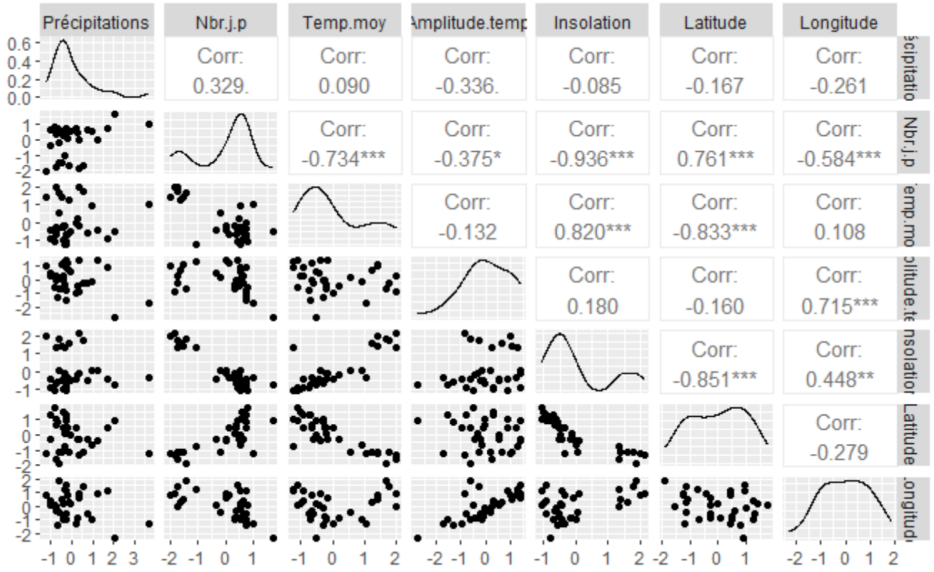
Notre objectif sera donc de construire un modèle linéaire afin de prédire le niveau des précipitations.

Choix du modèle

On va commencer par retirer les variables que l'on sait corrélés:

- On veut étudier les précipitations annuelles, on retire donc les précipitations mensuelles qui explique entièrement le modèle ainsi que les précipitations sommés de mai à août et de septembre à octobre
- On retire le nombre de jours de pluie mensuelle pour garder seulement le nombre de jours de pluie annuel
- La Latitude et la Longitude suffise pour décrire la localisation: on retire le nom des villes ainsi que "géographie"

On fait ensuite une vérification manuelle de la corrélation entre les différentes variables restantes afin d'avoir un bon modèle avant d'entamer l'affinement de celui-ci.



Plusieurs problèmes ressortent de graphique dont seules les valeurs numériques nous intéressent:

- Les variables semblent assez peu corrélés avec les précipitations, rendant difficile sa prédiction
- Les variables sembles très corrélés entre elles, ce qui complique l'analyse du modèle et peut fausser les résultats.

Nous allons poser 3 modèles pour essayer de gérer le problème des corrélations de façon différentes :

- **1er modèle:** On supprime toutes les variables fortements corrélés (avec *** sur le graph)
- **2e modèle:** On supprime les variables dont la valeur absolue de corrélation est au dessus de 0.8
- **3e modèle:** On ne supprime aucune variable

Pour choisir laquelle des deux variables corrélés supprimer, on choisira de de retirer celle avec la plus petite corrélation avec la variable d'intérêt "Précipitations".

Modèle 1

Elimination des variables corrélés

On retire une par une les variables les plus corrélés entre elles jusqu'a n'avoir que des corrélations "mineurs".

En appliquant la méthode expliqué précédemment on se retrouve avec deux variables: "Nombre de jours de pluie" et "Amplitude des températures" qui sont peu corrélés (comparés aux corrélations supprimés):

Deux variables n'est pas dutout un nombre suffisant pour un modèle aussi complexe, nous allons donc continuer l'analyse de ce modèle en faisant les techniques habituelles mais on sait déja que le résultat sera complètement faux.

Choix du modèle 1

On utilise un algorithme de Backward-Forward pour déterminer le meilleur modèle à partir de nos deux variables restantes, puis on fait une régréssion linéaire pour connaître la significativité du modèle final.

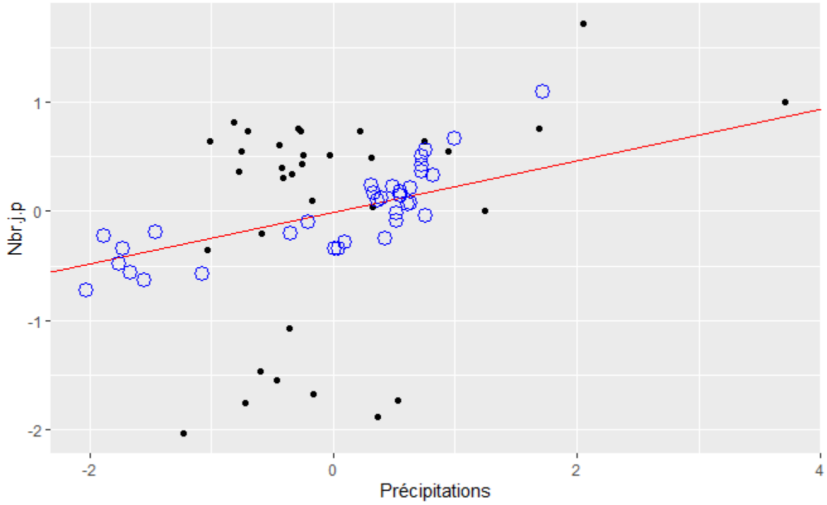
```
Start:  AIC=-0.99
Précipitations ~ Nbr.j.p + Amplitude.temp

                Df Sum of Sq    RSS      AIC
- Nbr.j.p        1    1.5849 29.269 -1.09440
<none>                      27.684 -0.98716
- Amplitude.temp 1    1.7385 29.422 -0.91641

Step:  AIC=-1.09
Précipitations ~ Amplitude.temp

                Df Sum of Sq    RSS      AIC
<none>                      29.269 -1.09440
+ Nbr.j.p        1    1.5849 27.684 -0.98716
- Amplitude.temp 1    3.7311 33.000  0.98500

Call:
lm(formula = Précipitations ~ Amplitude.temp, data = data1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.172e-18  1.644e-01   0.000   1.0000
Nbr.j.p     3.293e-01  1.669e-01   1.973   0.0572 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9589 on 32 degrees of freedom
Multiple R-squared:  0.1084,    Adjusted R-squared:  0.08055
F-statistic: 3.891 on 1 and 32 DF,  p-value: 0.05724
```

On se retrouve avec une unique variable assez peu significative, avec un modèle inutile puisqu'on a une valeur du R^2 de 0.11: c'est à dire que seulement 11% du modèle explique les précipitations.

D'après la P-valeur obtenue on peut même accepter l'hypothèse que les variables n'expliquent pas dutout le modèle.

Visualisation

Profitons du fait que l'on soit en dimension 1 pour tracer la droite des moindres carrés:



Nous voyons encore une fois que le modèle est non representatif: retirer toutes les variables corrélés n'est pas la meilleure solution.

Modèle 2

Elimination des variables corrélés On applique la même méthode mais seulement pour les variables dont la corrélation en valeur absolue est supérieure à 0.8 :

On se retrouve alors avec 4 variables explicatives (plus ou moins corrélés): "Nombre de jours de pluie", "Amplitude des températures", "Latitude" et "Longitude", ce qui semble un nombre correcte de variables pour avoir un modèle plutôt représentatif.

Choix du modèle 2 Nous réutilisons les deux mêmes fonctions :

```
Start:  AIC=-21.24
Précipitations ~ Nbr.j.p + Amplitude.temp + Latitude + Longitude

                Df Sum of Sq    RSS      AIC
<none>                      13.566 -21.2391
- Amplitude.temp  1    1.7259 15.292 -19.1673
- Longitude       1    1.9131 15.479 -18.7538
- Nbr.j.p         1   13.8198 27.386   0.6445
- Latitude        1   13.8601 27.426   0.6946

Call:
lm(formula = Précipitations ~ Nbr.j.p + Amplitude.temp + Latitude +
    Longitude, data = data2)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.702e-15  1.173e-01   0.000   1.0000
Nbr.j.p         1.244e+00  2.289e-01   5.435 7.57e-06 ***
Amplitude.temp -3.281e-01  1.708e-01  -1.921   0.0646 .
Latitude       -1.052e+00  1.932e-01  -5.443 7.40e-06 ***
Longitude       4.063e-01  2.009e-01   2.022   0.0525 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.684 on 29 degrees of freedom
Multiple R-squared:  0.5889,    Adjusted R-squared:  0.5322
F-statistic: 10.39 on 4 and 29 DF,  p-value: 2.408e-05
```

Nous gardons donc nos 4 variables, qui semblent toutes contenir au moins un peu d'information; et nous observons un R^2 qui semble correcte, avec une valeur de 59% (~5x plus que celui du 1er modèle).

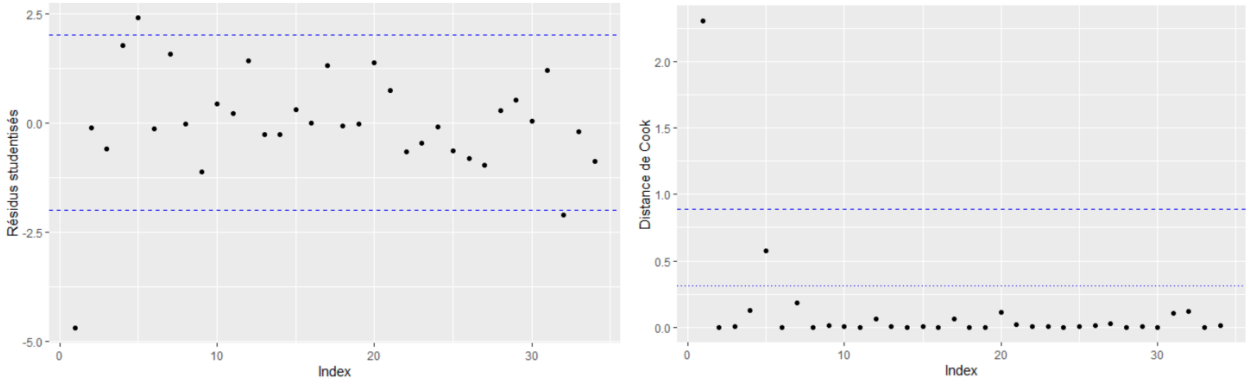
D'après la P-valeur obtenue on conclut que notre modèle contient au moins une variable qui explique la variable d'intérêt.

Ce second modèle semblant plutôt bon, on va faire sur celui-ci une étude des résidus.

Cependant il ne faut pas oublié que ce modèle parait cohérent avec une grande corrélation entre les variables, corrélations qui peuvent fausser les résultats: nous verrons via les prédictions ce qu'il en est vraiment.

Etude des résidus

Concentrons-nous sur les résidus studentisés : nous allons regarder si on a des valeurs abérrantes, puis regarder la distance de cook pour regarder l'impact de ces valeurs abérrantes sur le modèle (si existantes)

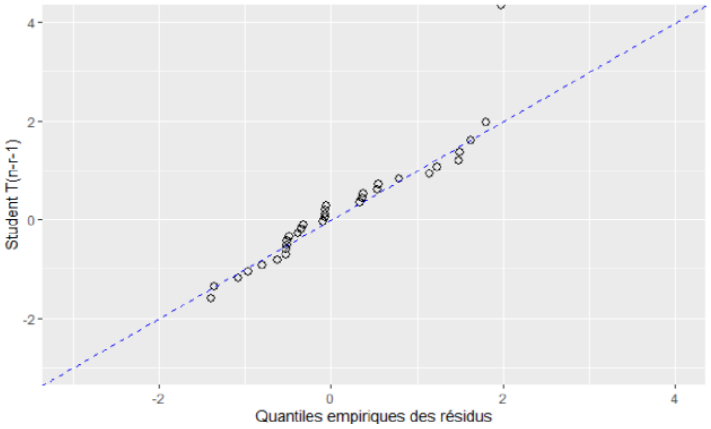


On a environ 11% des résidus valeurs en dehors de l'intervalle, soit le double de la valeur de référence à 5%. On a donc des valeurs abérrantes comme on pouvait s'en douter.

Concernant la distance de Cook nous observons qu'une seule valeur est vraiment abérrante: la 1ere. Or cette valeur correspond à celle de la ville d'Ajaccio.

Cette abération pourrait donc s'expliquer par le climat différent en Corse qu'en France métropolitaine ainsi que son éloignement géographique par rapport aux autres villes, on va donc décider d'exclure cette variable du modèle (ainsi que des modèles à venir).

Observons maintenant la répartition des résidus pour conclure sur leur normalité.




```
        Exact one-sample Kolmogorov-Smirnov test

data:  df.residu_stu2$residu_stu
D = 0.10358, p-value = 0.8227
alternative hypothesis: two-sided
```

On peut conclure d'après les quantiles empiriques des résidus studentisés et le test de Kolmogorov-Smirnov et que les résidus suivent une loi gausienne centrée de variance finit.

On va maintenant tester le modèle etc...

Prédiction du modèle

....

Modèle 3

Choix du modèle 3 Puisque dans ce modèle on conserve toutes les variables corrélés, on peut utiliser les méthodes Backward-Forward et de régréssion:

```
Start:  AIC=-44.71
Précipitations ~ Nbr.j.p + Temp.moy + Amplitude.temp + Insolation +
    Latitude + Longitude

                Df Sum of Sq     RSS     AIC
- Temp.moy       1    0.2961  5.8657 -45.004
<none>                       5.5697 -44.713
- Amplitude.temp  1    1.2824  6.8521 -39.875
- Insolation      1    1.8284  7.3981 -37.344
- Latitude        1    3.7365  9.3061 -29.772
- Longitude       1    5.3574 10.9270 -24.474
- Nbr.j.p         1   11.0832 16.6528 -10.570

Step:  AIC=-45
Précipitations ~ Nbr.j.p + Amplitude.temp + Insolation + Latitude +
    Longitude

                Df Sum of Sq     RSS     AIC
<none>                       5.8657 -45.004
+ Temp.moy       1    0.2961  5.5697 -44.713
- Insolation      1    1.7245  7.5902 -38.498
- Amplitude.temp  1    2.4226  8.2883 -35.595
- Longitude       1    5.2026 11.0683 -26.050
- Latitude        1    7.2388 13.1045 -20.477
- Nbr.j.p         1   11.5006 17.3663 -11.185

Call:
lm(formula = Précipitations ~ Nbr.j.p + Amplitude.temp + Insolation +
    Latitude + Longitude, data = data_v2)
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.09541    0.08215   1.161  0.25563
Nbr.j.p          2.18663    0.30054   7.276 7.96e-08 ***
Amplitude.temp  -0.53110    0.15904  -3.339  0.00246 **
Insolation       0.97744    0.34693   2.817  0.00894 **
Latitude        -1.03134    0.17867  -5.772 3.86e-06 ***
Longitude        0.85624    0.17497   4.894 4.05e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.4661 on 27 degrees of freedom
Multiple R-squared:  0.8202,    Adjusted R-squared:  0.7869
F-statistic: 24.64 on 5 and 27 DF,  p-value: 2.814e-09
```

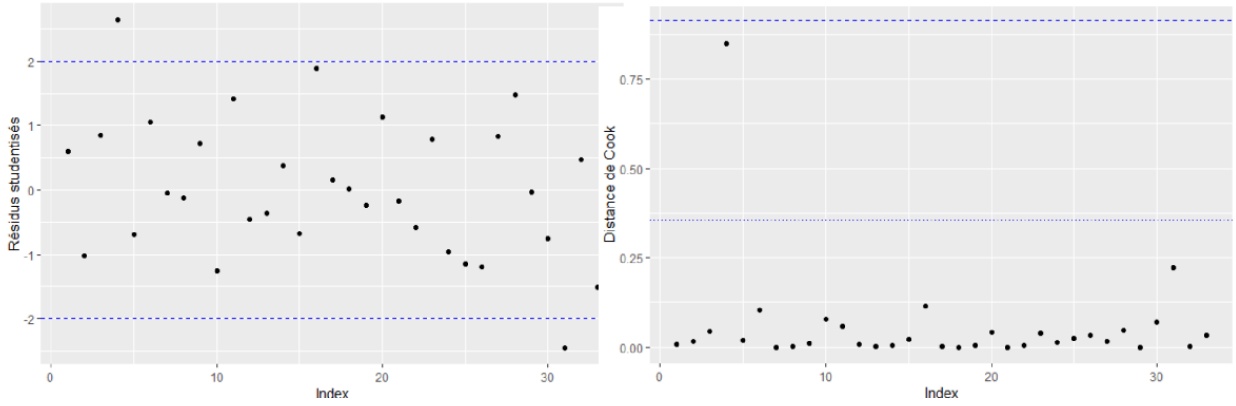
On a donc a 5 variables pour expliquer notre modèle (plus grand nombre jusqu'à présent) avec toutes de la significativité.

De plus on observe le meilleur R^2 jusqu'à maintenant d'une valeur de 82% (même en prenant le R^2_a qui pénalise le nombre de variables on est toujours à 20% de plus que le R^2 du modèle 2).

Nous rappelons encore une fois que ce modèle paraît certe le plus représentatif mais est aussi le plus corrélé.

Pour comparer véritablement ce modèle au précédent nous allons faire également une analyse des résidus et des prédictions.

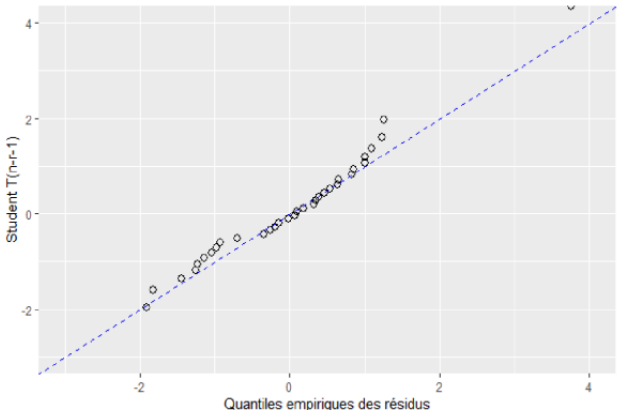
Etude des résidus Concentrons-nous encore une fois sur les résidus studentisés avec les potentiels valeurs abérrantes, et leur distance de cook pour regarder l'impact de ces valeurs abérrantes sur le modèle (si existantes):



On observe 6% de valeurs en dehors de l'intervalle, ce qui est proche de la limite de 5% et il paraît donc acceptable de dire qu'il n'y a pas de données abérrantes. Cependant étant donnée la forte corrélation des données on a préféré tracé aussi les distances de Cook:

Un point dépasse le 1er seuil mais aucun ne dépasse second qui est problématique, on peut donc garder le modèle tel quel.

Concernant la normalité des résidus, on confirme comme pour le modèle 2 leur gaussianité.




```
        Exact one-sample Kolmogorov-Smirnov test

data:  df.residu_stu3$residu_stu
D = 0.071334, p-value = 0.9917
alternative hypothesis: two-sided
```

Prédiction du modèle .....