Partie 1 Projet MRR

DURAND Lénaïc, SHI DE MILLEVILLE Guillaume

Binôme n°1, dataset Behavior of the urban traffic of the city of Sao Paulo in Brazil

Première feuille

Nature du problème

Ce projet est porté sur **l'étude du trafic urbain à Sao Paulo, au Brésil**. Avec l'avancement technologique des dernières années, de nombreuses problématiques touchant les transports et la logistique émergent. Le but est de **trouver un modèle de régression permettant de prédire la lenteur du trafic** en fonction des paramètres fournis par la base de données.

La **variable cible est donc *Slowness in traffic***, nous allons étudier ses variations en fonction des paramètres les plus influents.

Paramètres

Les paramètres sont les suivants :

- ***Hour*** l'heure à laquelle les données sont relevées : arrondie à la demi-heure près, elle prend des valeurs entre 7h et 20h, et les relevés sont faits du Lundi au Vendredi inclus.
- ***Immobilized_bus*** le nombre de bus immobilisés
- ***Broken_Truck*** le nombre de camions accidentés
- ***Vehicle_excess*** le surplus de véhicules
- ***Accident_victim*** le nombre de victimes d'accidents
- ***Running_over*** le nombre de personnes renversées par un véhicule
- ***Fire_vehicle*** le nombre de véhicules de pompier
- ***Occurence_involving_freight*** le nombre d'accidents impliquant une cargaison
- ***Incident_involving dangerous_freight*** le nombre d'incidents impliquant une cargaison dangereuse
- ***Lack_of_electricity*** le niveau d'importance du manque d'électricité
- ***Fire*** la présence ou non d'un incendie
- ***Point_of_flooding*** le nombre de sources d'inondation
- ***Manifestations*** la présence ou non de manifestations
- ***Defect_in_the_network_of_trolleybuses*** le nombre de défauts dans le réseau de tramways
- ***Tree_on_the_road*** la présence ou non d'arbre(s) sur la route
- ***Semaphore_off*** le nombre de feux de circulation ne fonctionnant pas
- ***Intermittent_semaphore*** la présence ou non de feux de circulation temporaires

Influence des paramètres

Afin d'exploiter au mieux les données, on calcule la moyenne de la variable de *Slowness_in_traffic* pour toutes les autres variables afin d'en observer une tendance. Une telle manipulation permet de **souligner l'augmentation de *Slowness_in_traffic* aux heures d'affluences**. Ces calculs permettent uniquement d'obtenir des observations préliminaires sur les variables. Ainsi, les données qui seront utilisées par la suite seront toujours les données brutes.

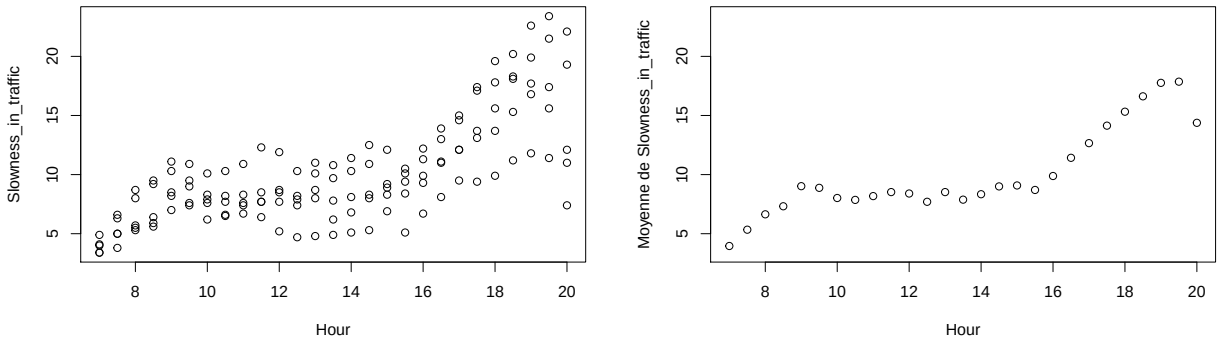
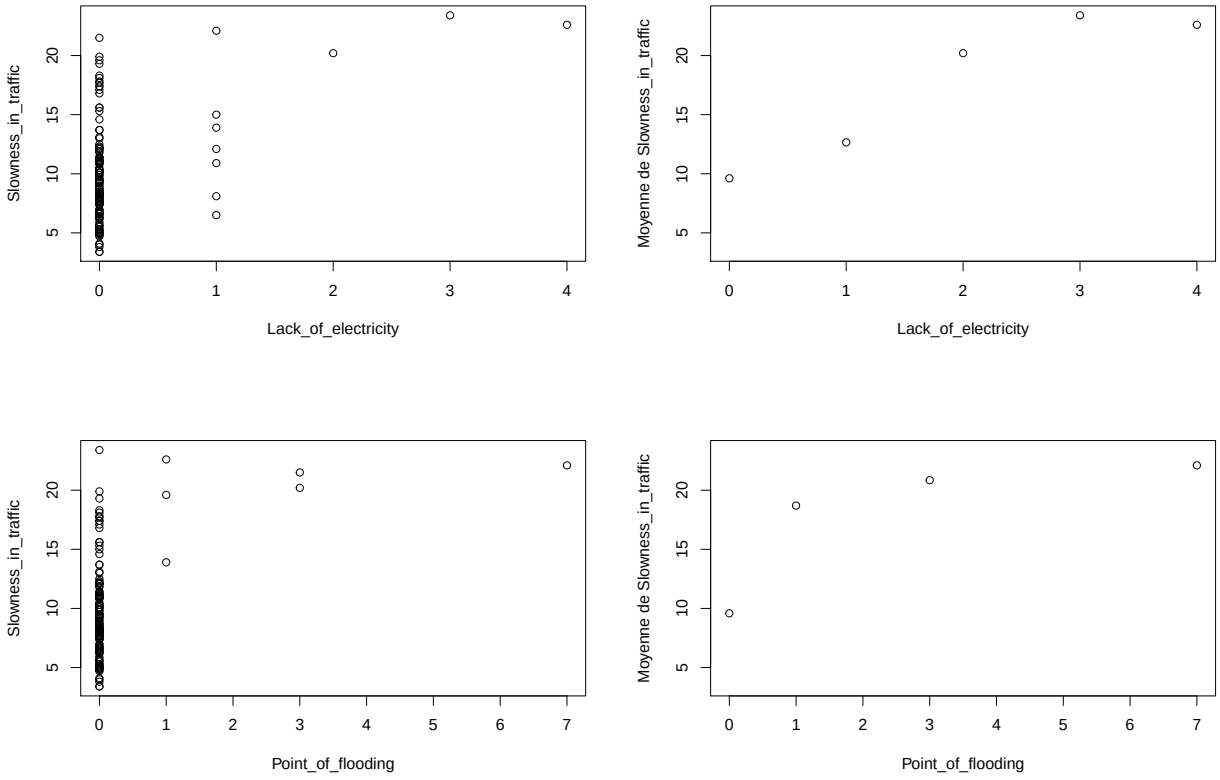


Figure 1 : *influence de Hour sur Slowness_in_traffic*

On voit bien qu'aux heures d'affluence il y a une augmentation de la variable *Slowness_in_traffic* puis un pic (augmentation de la variable jusqu'aux pics a 9h et 19h30). **Un tel comportement est attendu** puisqu'il correspond à la réalité.

D'autres variables telle que *Semaphore off* ont également un comportement intéressant :



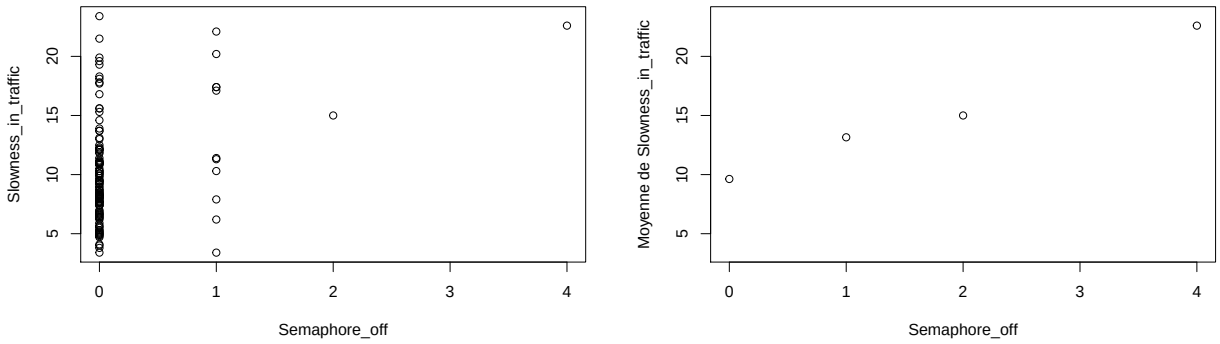


Figure 2 : influence de Semaphore_off_sur Slowness_in_traffic

On observe alors un **comportement linéaire** dans le résumé obtenu, tout comme pour la variable *Lack of electricity*.

Certaines variables comme *Incident involving dangerous freight* ou *Occurence involving freight* **ne contiennent presque que des 0** et ne semblent pas avoir d'influence sur *Slowness in traffic*. Pour les mêmes raison on retire également *Intermittent Semaphore* et *Fire vehicle*.

Manifestaton contient bien moins de 0 mais **ne semble pas influencer *Slowness in traffic*** puisque la variable cible n'évolue pas pour un nombre grandissant de manifestations. On élimine donc également *Defect in the network of trolley buses*, *Trees on the road*, *Immobilized bus*, *Broken truck*, *Vehicle excess*, *Accident victim*, *Running over* et *Fire*.

L'étude se portera donc sur l'influence de ***Hour***, ***Lack of electricity***, ***Point of flooding*** et ***Semaphore off*** sur ***Slowness in traffic***.

Seconde feuille

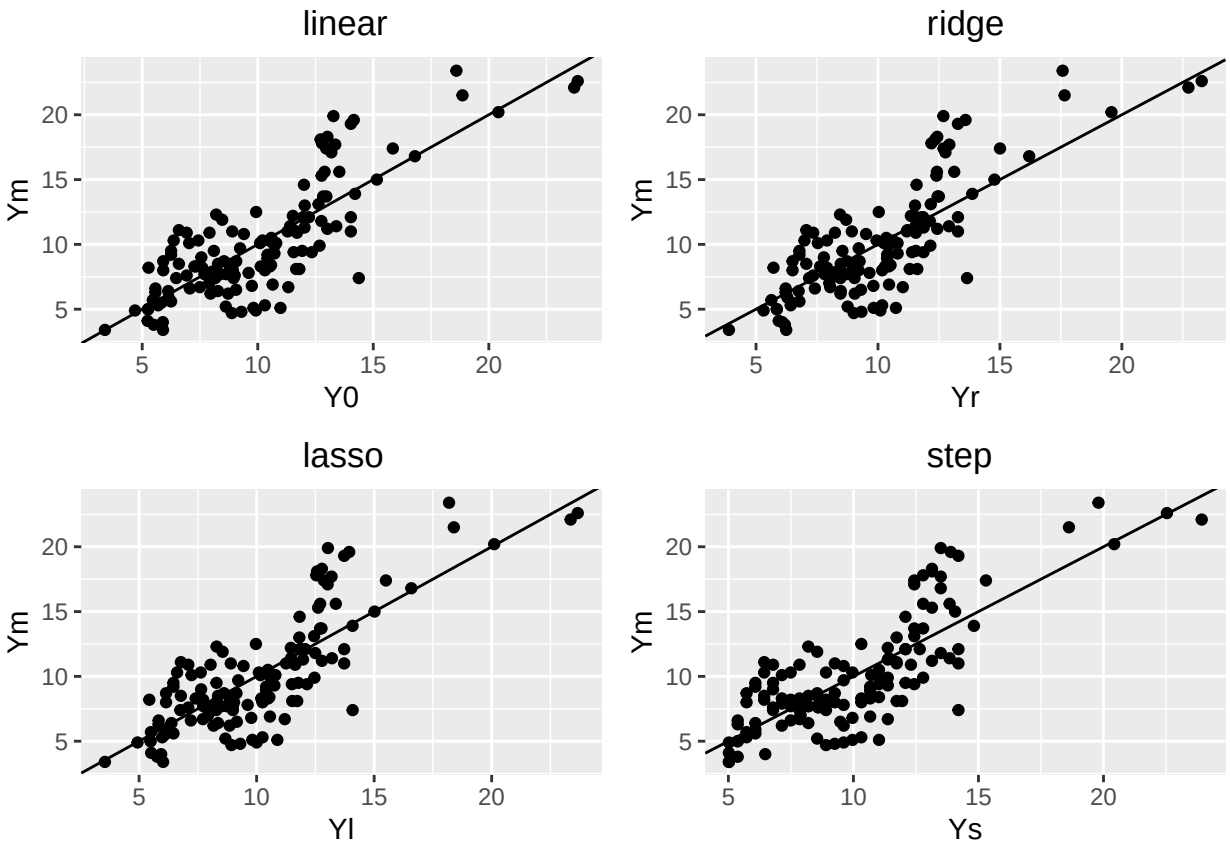
Introduction

Dans la partie précédente, nous avons présenté le problème et les données utilisées. Nous avions aussi présenté les variables explicatives sur lesquelles nous souhaitions travailler : ***Hour***, ***Lack of electricity***, ***Point of flooding*** et ***Semaphore off***. Le but de cette partie est d'expliquer la méthodologie suivie pour trouver le meilleur modèle possible.

Tout d'abord, nous allons comparer les variables explicatives obtenues avec un algorithme de sélection à celles que nous avions sélectionnées manuellement. Par exemple avec la méthode **stepwise**, nous trouvons les variables ***Hour***, ***Lack of electricity***, ***Point of flooding*** et ***Manifestations***, ce qui correspond a 75% à notre analyse en amont. Ensuite, nous présenterons un second modèle de classification.

Régression

Nous commençons par regarder les résultats donnés par d'autres types de modèles sélectifs, tels que ridge et lasso. Pour cela nous effectuons d'abord une validation croisée pour trouver le paramètre λ optimal, c'est-à-dire celui qui minimise l'erreur.



Les R^2 correspondants aux graphes ci-dessus sont les suivants :

```
##   Linear_R2  Ridge_R2  Lasso_R2   Step_R2
## 1  0.657544 0.6552706 0.6571681 0.6427598
```

La RMSE nous donne plus d'information sur l'efficacité de ces modèles. Le R^2 varie beaucoup en fonction de la variance et ne peut être interprétée que difficilement.

```
##   Linear_RMSE Ridge_RMSE Lasso_RMSE Step_RMSE
## 1    2.543884   2.581076   2.549745  2.598215
```

On observe que les résultats des modèles sont assez similaires : un R^2 d'environ 0,65 et une RMSE d'environ 2,6 (ce qui correspond à une **erreur d'environ 10%** pour la valeur maximale de *Slowness in trffic* 23,4). Ce sont des **résultats satisfaisants**. Il est cependant surprennant de voir que le meilleur modèle est le modèle de base, c'est-à-dire un modèle linéaire prenant en compte toutes les variables. Or, la première étude proposait plutôt de ne garder que quelques variables. Afin d'obtenir plus d'information sur ces données, on se propose d'étudier un autre modèle.

Classification

Cet modèle possible consiste à créer une nouvelle variable binaire Y tel que $Y = \mathbb{1}_{Slowness\ in\ traffic\ >\ 3^{ème}\ quartile}$, et faire une régression logistique. Ainsi, on cherche un modèle capable de prédire s'il faut prendre sa voiture ou non. On obtient ainsi un **modèle de classification permettant de dire si le trafic est mauvais** (considéré mauvais si supérieur au $3^{ème}$ quartile.)

On obtient alors la matrice de confusion suivante :

```
##          Ylog
## Y_binaire  0  1
##         0 96  5
##         1 10 24
```

Ayant obtenu le nombre de faux positifs, faux négatifs, vrais positifs et vrais négatifs, il est aisé d'en déduire les critères d'évaluation suivants :

```
##      recall precision  F1_score
## 1 0.8275862 0.7058824 0.7619048
```

Conclusion

Afin **d'évaluer les performances** des modèles de prédiction, nous utiliserons la méthode du **K-fold**. Les perfomances seront **évaluées à l'aide du coefficient de régression et de la somme des résidus**. Au final nous choisirons le modèle donnant le meilleur résultat. Ce modèle servira à prédire avec précision la densité du trafic.

En parallèle, nous disposerons également de notre modèle de classification permettant de dire si le trafic sera très mauvais ou non.

Préparation de la soutenance

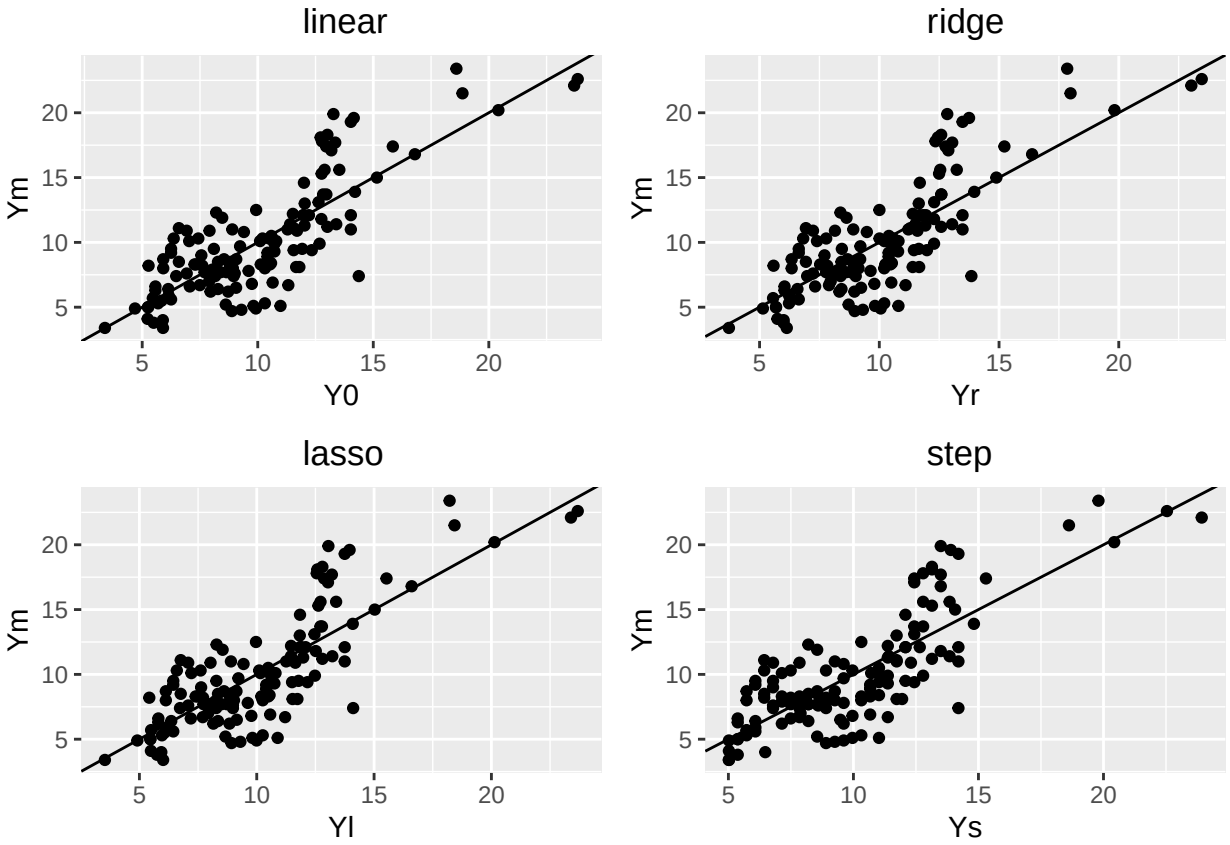
Introduction

Dans la partie précédente, nous avons présenté le problème et les données utilisées. Nous avions aussi présenté les variables explicatives sur lesquelles nous souhaitions travailler : ***Hour***, ***Lack of electricity***, ***Point of flooding*** et ***Semaphore off***. Le but de cette partie est d'expliquer la méthodologie suivie pour trouver le meilleur modèle possible.

Tout d'abord, nous allons comparer les variables explicatives obtenues avec un algorithme de sélection à celles que nous avions sélectionnées manuellement. Par exemple avec la méthode **stepwise**, nous trouvons les variables ***Hour***, ***Lack of electricity***, ***Point of flooding*** et ***Manifestations***, ce qui correspond a 75% à notre analyse en amont. Ensuite, nous présenterons un second modèle de classification.

Régression

Nous commençons par regarder les résultats donnés par d'autres types de modèles sélectifs, tels que ridge et lasso. Pour cela nous effectuons d'abord une validation croisée pour trouver le paramètre λ optimal, c'est-à-dire celui qui minimise l'erreur.



Les R^2 correspondants aux graphes ci-dessus sont les suivants :

```
##   Linear_R2 Ridge_R2  Lasso_R2   Step_R2
## 1  0.657544  0.65631 0.6572271 0.6427598
```

La RMSE nous donne plus d'information sur l'efficacité de ces modèles. Le R^2 varie beaucoup en fonction de la variance et ne peut être interprétée que difficilement.

```
##   Linear_RMSE Ridge_RMSE Lasso_RMSE Step_RMSE
## 1    2.543884   2.563653   2.548814  2.598215
```

On observe que les résultats des modèles sont assez similaires : un R^2 d'environ 0,65 et une RMSE d'environ 2,6 (ce qui correspond à une **erreur d'environ 10%** pour la valeur maximale de *Slowness in trffic* 23,4). Ce sont des **résultats satisfaisants**. Il est cependant surprennant de voir que le meilleur modèle est le modèle de base, c'est-à-dire un modèle linéaire prenant en compte toutes les variables. Or, la première étude proposait plutôt de ne garder que quelques variables. Afin d'obtenir plus d'information sur ces données, on se propose d'étudier un autre modèle.

Classification

Cet modèle possible consiste à créer une nouvelle variable binaire Y tel que $Y = \mathbb{1}_{Slowness\ in\ traffic\ >\ 3^{ème}\ quartile}$, et faire une régression logistique. Ainsi, on cherche un modèle capable de prédire s'il faut prendre sa voiture ou non. On obtient ainsi un **modèle de classification permettant de dire si le trafic est mauvais** (considéré mauvais si supérieur au $3^{ème}$ quartile.)

On obtient alors la matrice de confusion suivante :

```
##          Ylog
## Y_binaire  0  1
##         0 96  5
##         1 10 24
```

Ayant obtenu le nombre de faux positifs, faux négatifs, vrais positifs et vrais négatifs, il est aisé d'en déduire les critères d'évaluation suivants :

```
##      recall precision  F1_score
## 1 0.8275862 0.7058824 0.7619048
```

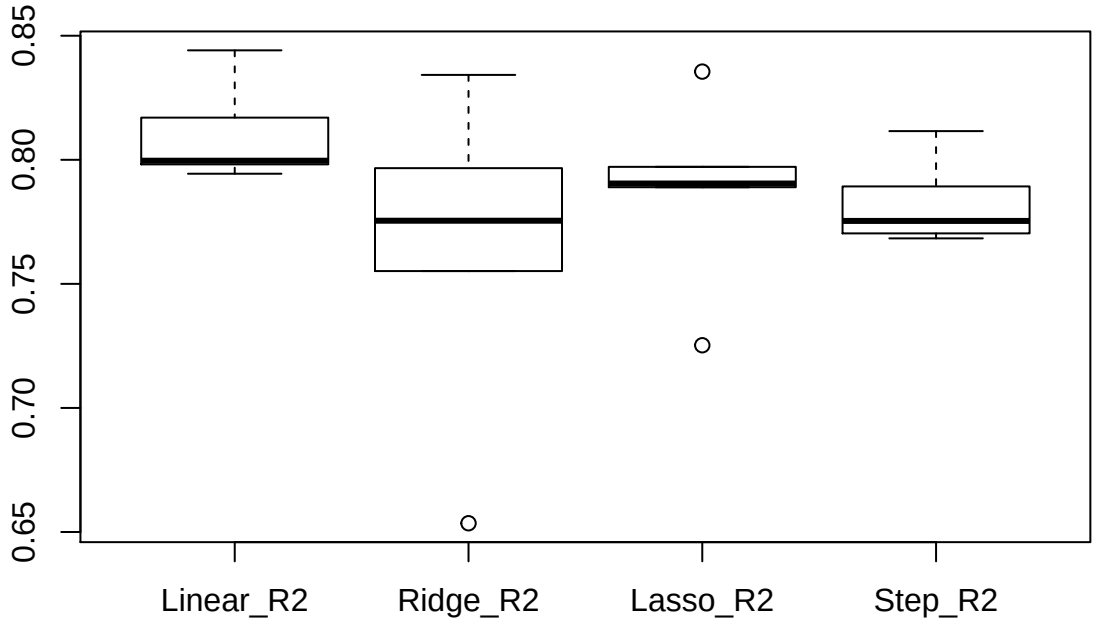
Conclusion

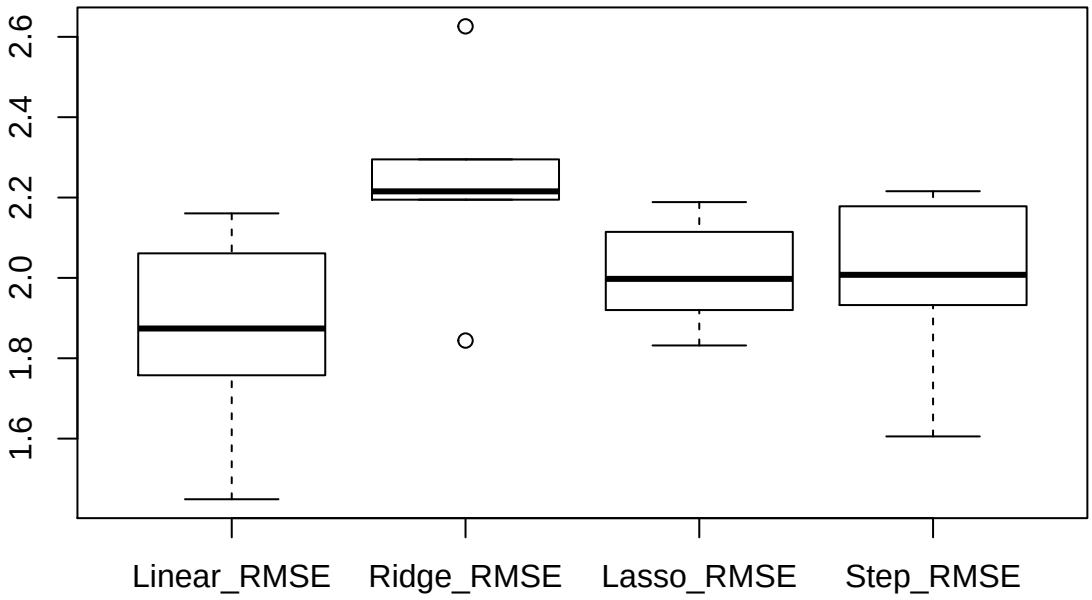
Afin **d'évaluer les performances** des modèles de prédiction, nous utiliserons la méthode du **K-fold**. Les perfomances seront **évaluées à l'aide du coefficient de régression et de la somme des résidus**. Au final nous choisirons le modèle donnant le meilleur résultat. Ce modèle servira à prédire avec précision la densité du trafic.

En parallèle, nous disposerons également de notre modèle de classification permettant de dire si le trafic sera très mauvais ou non.

```
##   moy_Linear_R2 moy_Ridge_R2 moy_Lasso_R2 moy_Step_R2
## 1     0.8106573     0.762999    0.7874709   0.7829778
```

```
##   moy_Linear_RMSE moy_Ridge_RMSE moy_Lasso_RMSE moy_Step_RMSE
## 1        1.860472       2.235062       2.010522      1.987902
```




1.6
1.8
2.0
2.2
2.4
2.6
Linear_RMSE
Ridge_RMSE
Lasso_RMSE
Step_RMSE