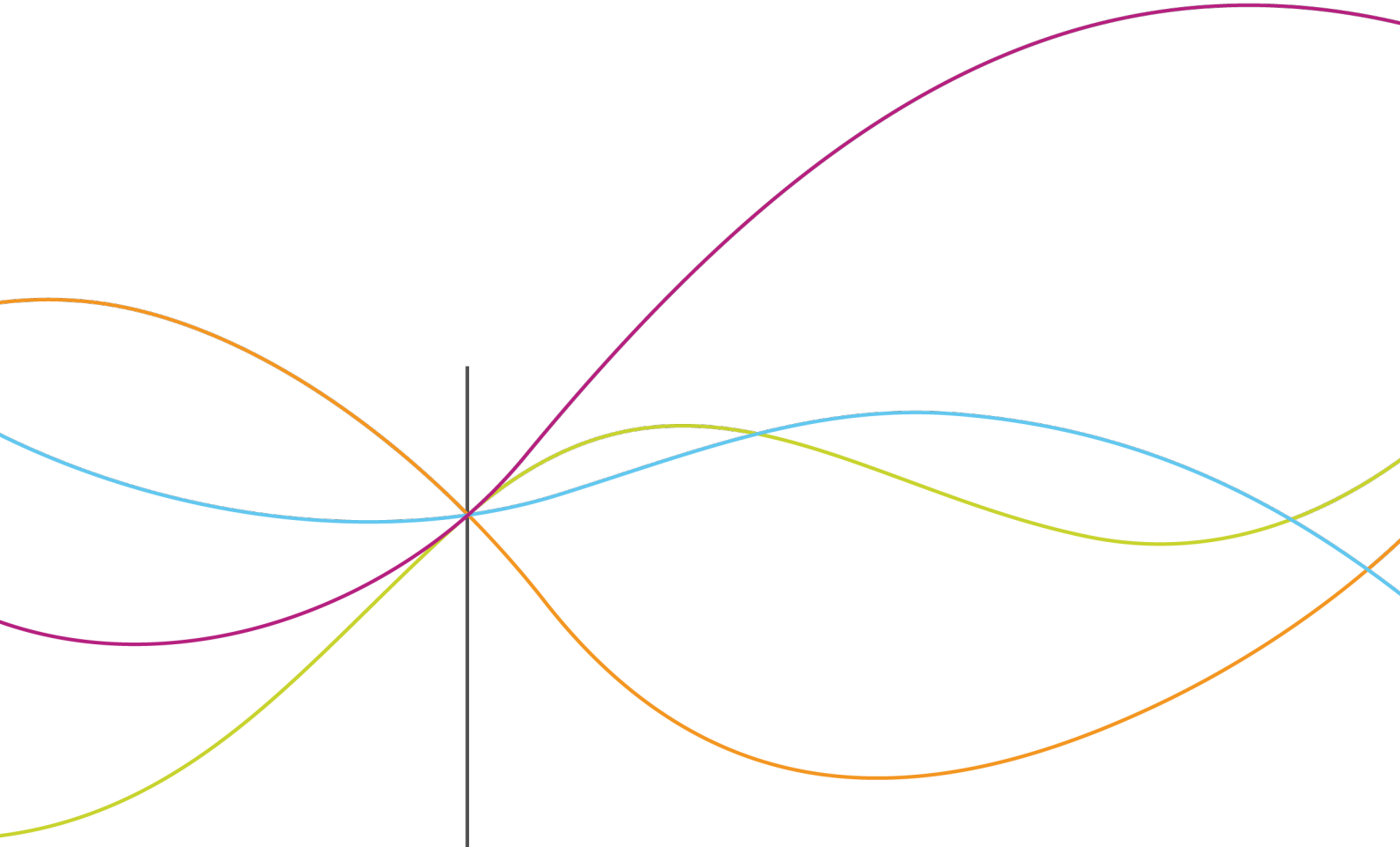
ENSAE ParisTech
École nationale de la statistique et de l'administration économique
université PARIS-SACLAY



Gouysse Margaux
Grelety Antoine
Watrigant Timothée

TP1 : Prédiction de l'efflorescence algale

2017/2018

Questions de la partie 4.6

- **(a).** La commande lm de R va transformer chaque variable catégorielle (season, size, speed) par des variables dites "dummy". Chaque catégorie devient une nouvelle variable qui prend 1 ou 0 pour chaque observation. R va alors supprimer une des variables binaires créées afin d'éviter la colinéarité.
- **(a).** On utilise le R^2 ajusté qui permet de mesurer la qualité d'ajustement par un modèle linéaire. Ici, $R^2 = 0.3204$. Plus R^2 est proche de 1, mieux c'est. On peut donc conclure ici que le résultat n'est pas trés bon car plus proche de 0 que de 1.
- **(b).** Les variables clairement inutiles pour prévoir la variable a1 grace à ANOVA sont : season, NH4, CHla.
- **(c).** Les variables retenues par le modèle final sont : mn02, mxPH, NH4, size, NO3, PO4.
- **(c).** La qualité d'ajustement a augmenté par rapport au modèle initial contenant toutes les variables. En effet, a1 s'explique mieux car le R^2 ajusté a augmenté méme si le R^2 lui, a diminué (ce qui est normal puisque nous avons moins de variables, c'est pourquoi nous utilisons le R^2 ajusté).

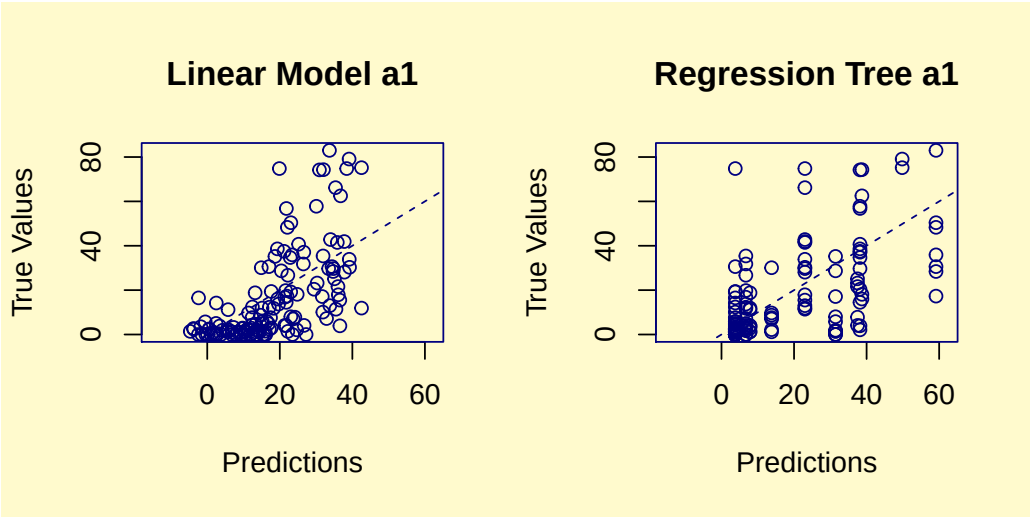
Prévisions de a1 sur test.algae : Commandes et résultats

Commandes

```
> test.algae = knnImputation(test.algae, k =10, meth = "median")
> lm.predictions.a1 = predict(final.lm,test.algae)
> rt.predictions.a1 = predict(rt.a1, test.algae)
```

On nettoie les données manquantes en remplaçant les valeurs "NA" de la table test.algae par la médiane de ses 10 plus proches voisins. On prédit ensuite la variable a1 de test.algae à l'aide des modèles prédictifs (régression linéaire multiple et arbre de décision) obtenus à partir de la table algae.

Résultats



Ici,la ligne en pointillés représente le cas idéal où la prédiction est égale à la réalité. Dans le cas de l'algue a1, on voit que l'on a un MAE plus petit pour l'arbre de décision, les points prédits sont donc plus proches en moyenne de la réalité. Cependant, le RMSE de l'arbre est plus grand que celui du modèle linéaire, ce qui souligne le plus grand nombre de valeurs aberrantes de l'arbre.

Prévisions de l'efflorence des algues a2, a3, a4, a5,a6, a7

Exemple du code pour une algue (a4)

```
> lm.a4 <- lm(a4 ~ ., data = algae[, c(1:11,15)])
> final.lma4 = step(lm.a4)
> rt.a4 = rpart(a4 ~ ., data = algae[, c(1:11,15)])
```

```
> lm.predictions.a4 = predict(final.lma4, test.algae)
> rt.predictions.a4 = predict(rt.a4, test.algae)
```

Pour chacune des algues, il faut refaire les deux modèles de prédiction (linéaire et arbre de décision) en veillant à ne pas prendre en compte les autres algues que celle qu'on cherche à prédire.

Résultats : Algue a2

```
       mae         mse        rmse        mape        nmse        nmae
 7.0854747 117.7113601  10.8494866         Inf   1.0963262   0.9255206

       mae         mse        rmse        mape        nmse        nmae
 6.8571224 102.6687228  10.1325576         Inf   0.9562238   0.8956927
```

Avec a2 commme variable de prédiction, le MSE et MAE du modèle linéaire sont plus faibles que ceux de l'arbre de décision. On observe toujours un biais positif pour les prédictions des valeurs faibles de a2 dans le modèle linéaire. Les prédictions de l'arbre de décision sont globalement plus éclatées autour de leur vraie valeur par rapport au modèle linéaire.

Résultats : Algue a3

```
      mae        mse       rmse       mape       nmse       nmae
4.3901411 40.6447811  6.3753260        Inf  1.2867263  0.9772886

      mae        mse       rmse       mape       nmse       nmae
3.8578505 28.2747627  5.3174019        Inf  0.8951181  0.8587955
```

Les métriques d'erreur indiquent que le modèle linéaire est plus approprié. Les graphiques confirment cette hypothèse. En particulier, l'arbre de décision prédit des valeurs trés biaisées pour a3 proches de zéro. Leur vraie valeur est significativement plus élevée (supérieure à 5).

Résultats : Algue a4

```
     mae       mse      rmse      mape      nmse      nmae
1.7969618 8.4343098 2.9041883       Inf 1.0743646 0.9081953

     mae       mse      rmse      mape      nmse      nmae
1.8803466 7.6807532 2.7714172       Inf 0.9783764 0.9503385
```

Les deux modèles prédisent les valeurs de a4 avec une précision équivalente. Les valeurs de a4 sont concentrées entre 0 et 1O. Les valeurs prédites sont relativement proches des vraies valeurs. Il n'y a pas de biais apparent pour chacun des modèles.

Résultats : Algue a5

```
      mae        mse       rmse       mape       nmse       nmae
5.2334536 90.2131493  9.4980603        Inf  0.9769874  0.8571152

      mae        mse       rmse       mape       nmse       nmae
5.4354391 80.6201197  8.9788707        Inf  0.8730971  0.8901956
```

Les deux modèles sont évalués avec une qualité de prédiction équivalente. Le modèle linéaire semble prédire a5 avec un léger biais positif pour les valeurs faibles. L'arbre de décision prédit a5 avec une volatilité plus forte pour les valeurs plus élevées.

Résultats : Algue a6

```
       mae         mse        rmse        mape        nmse        nmae
 6.8209822 140.3108602  11.8452885         Inf   0.7821130   0.8050039

       mae         mse        rmse        mape        nmse        nmae
 7.2199867 145.6010046  12.0665241         Inf   0.8116010   0.8520939
```

Les résultats de prédiction de a6 sont similaires à a5. On remarque que pour les deux modèles, les valeurs prédites de a6 supérieures à 20 sont plus dispersées par rapport à leur vraire valeur.

Résultats : Algue a7

```
      mae        mse       rmse       mape       nmse       nmae
2.4819099 22.6205324  4.7561047        Inf  1.0452136  0.9228783

     mae       mse      rmse      mape      nmse      nmae
2.902611 24.004400  4.899428       Inf  1.109157  1.079313
```

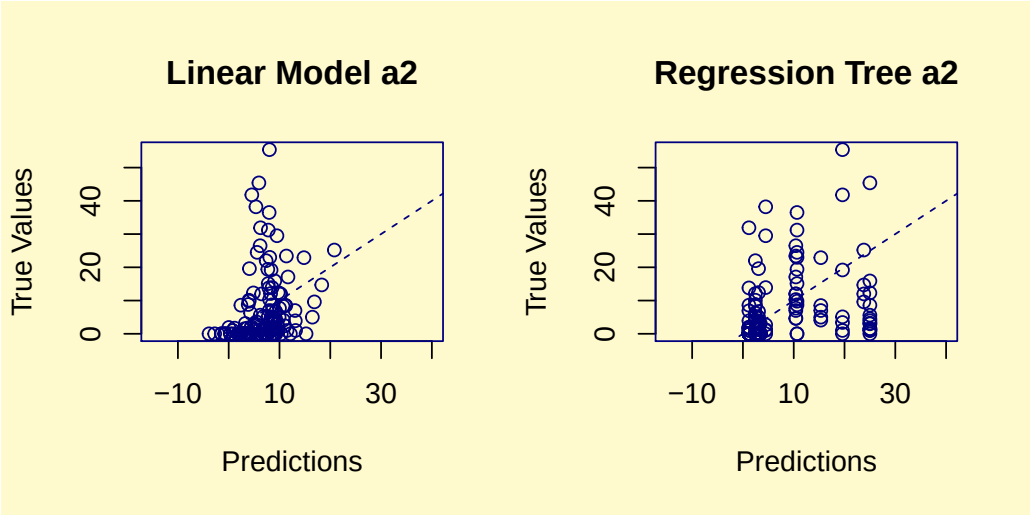
Pour a7, la qualité de prédiction de l'arbre décision est supérieure au modèle linéaire. La plupart des vraies valeurs de a7 sont concentrées entre 0 et 4. On observe de nouveau un biais positif pour le modèle linéaire, alors que les valeurs prédites pour l'arbre de décision sont réparties également autour de leur vraie valeur.

Conclusion

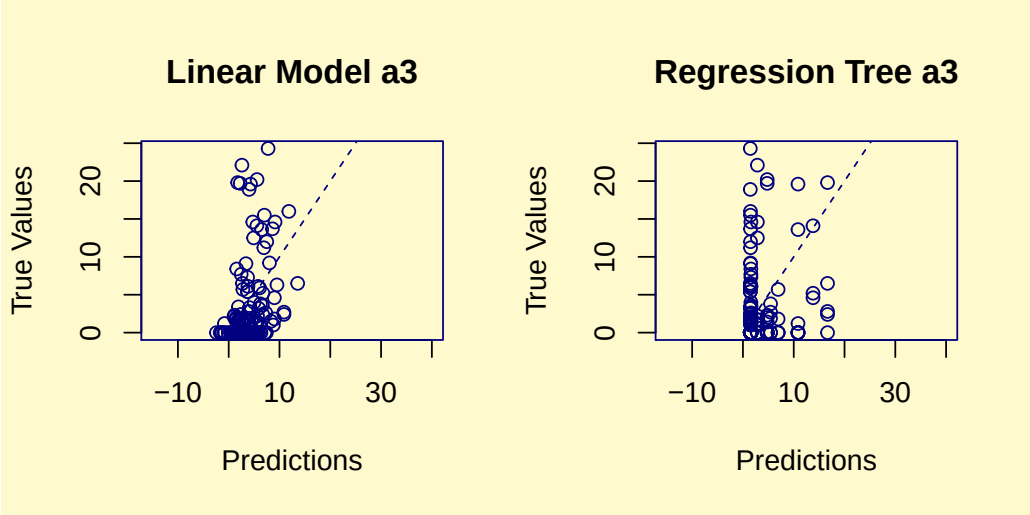
Sur les différentes observations on voit que l'on a un biais positif pour les valeurs basses dans le modèle linéaire, le modèle a tendance à surestimer la concentration d'algues lorsque leur vraie valeur est dans la tranche basse. La préférence pour un modèle dépend souvent de la métrique d'erreur considée (MAE ou RMSE). Le RMSE a tendance à attribuer plus de poids aux valeurs aberrantes, tandis que la MAE considère uniquement la distance moyenne par rapport à la vraie valeur. On constate également que le modèle linéaire peut prédire des valeurs négatives, ce qui peut être dérangeant dans certains cas, contrairement à l'arbre de décision qui lui ne prédit que des valeurs positives. Sur les différents cas, on voit que l'on a des modèles globalement insatisfaisants (que ce soit le modèle linéaire ou l'arbre de décision), puisque l'on a trop d'incertitudes sur la prédiction par rapport à la réalité. Il faudrait mettre en place un autre type de modèle, par exemple de type quadratique, puisque nos modèles ont tendance à prédire des valeurs plus grandes que la réalité pour les valeurs faibles et inversement pour les valeurs importantes.

A Annexes

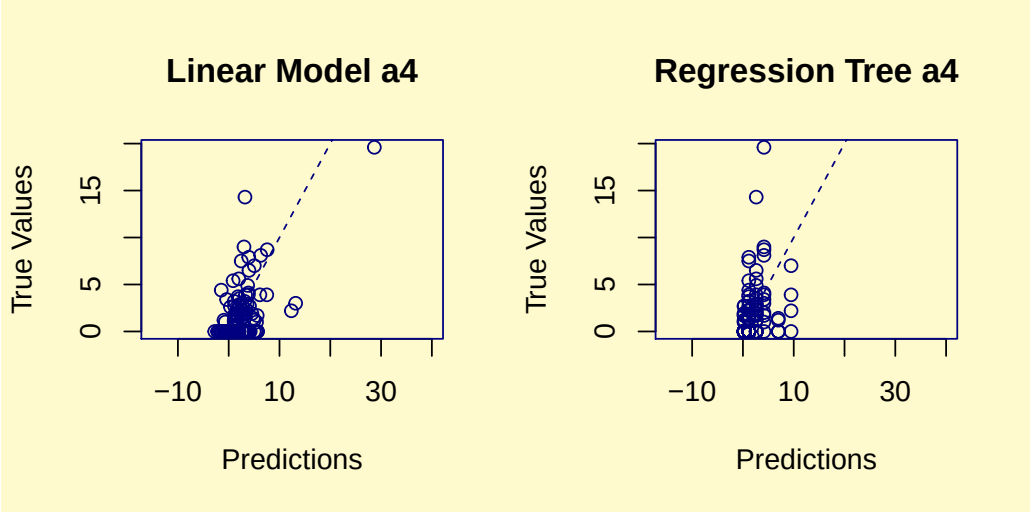
Algue a2



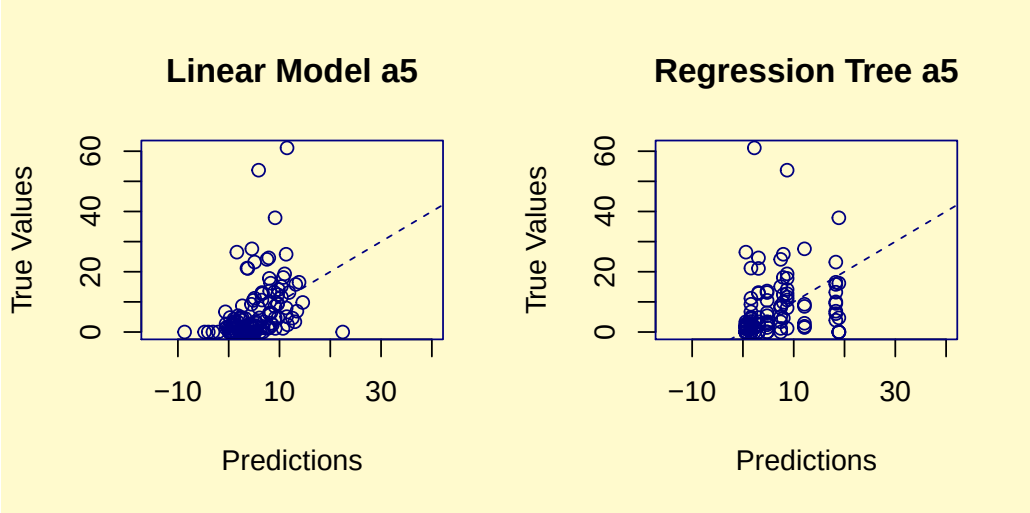
Algue a3



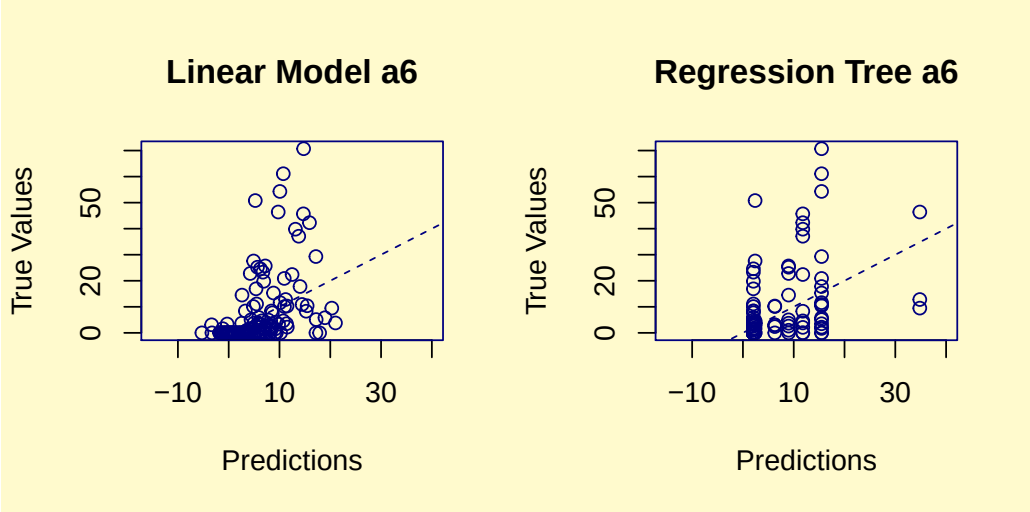
Algue a4



Algue a5



Algue a6



Algue a7