Generalized additive Model - GAM

Marie-Pierre Etienne based on F. Husson's slides

https://github.com/MarieEtienne

Novembre 2018

Outline

Outline

Le problème et les données

- Ozone phénomène complexe
- Enjeux de santé publique
- Mission Air Breizh : mesure, analyse, prévision $\rightarrow$ envoie tous les jours à 17 heures, l'indice de pollution du lendemain aux autorités

Le problème et les données

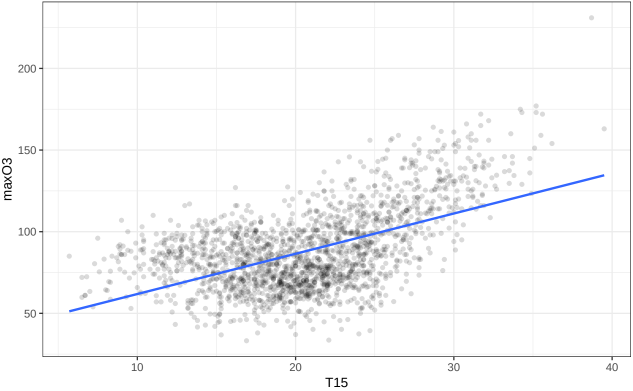
Prévoir les pics d'ozone en fonction des prévisions météorologiques à Rennes (Air Breizh)

	maxO3	**T6**	**T9**	**T12**	**T15**	**T18**	**Ne6**	**...**	**maxO3v**
19940401	**56**	8.6	9.5	6.8	9.1	7.7	6		59.6
19940402	**39.2**	3.6	5.6	9.2	8.4	4.9	3		56
19940403	**36**	2.7	7.3	6.3	7	7.9	6		39.2
19940404	**41.2**	11.8	11.8	11	7	7.7	8		36
19940405	**27.6**	3.7	8.3	11.6	10.7	7.9	6		41.2
	...								...
	...								...
20050929	**73**	11.2	16	17.8	18.6	15.1	2		68
20050930	**46**	14.2	17.3	17.2	17.5	18	8		73

Figure 1: Extrait données Ozone

La régression linéaire simple

```
p <- ggplot(data=ozone, aes(x = T15, y = maxO3)) + geom_point(alpha = 0.15)
p + geom_smooth(method = lm, se = FALSE)
```



$$Y_i = f(x_i) + E_i$$
$$f(x_i) = \beta_0 + \beta_1 x_i$$

Outline

Approche paramétrique

- fonction de régression connue
- dépend d'un certain nombre de paramètres
- paramètres estimés à partir des données
- attractif car interprétation des paramètres et simplicité statistique

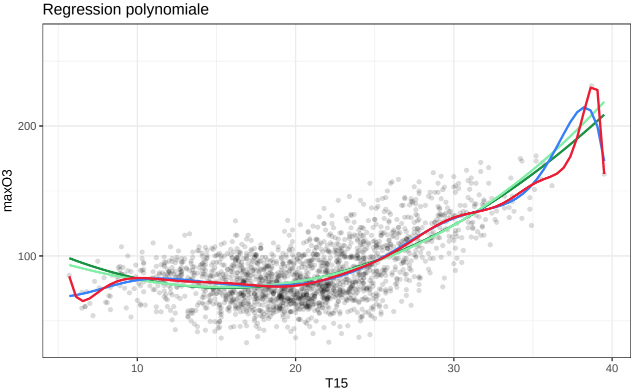
Exemple le plus simple: la régression linéaire simple

Mais ne reflète pas toujours la relation entre Y et x.

Approche paramétrique: régression linéaire polynomiale

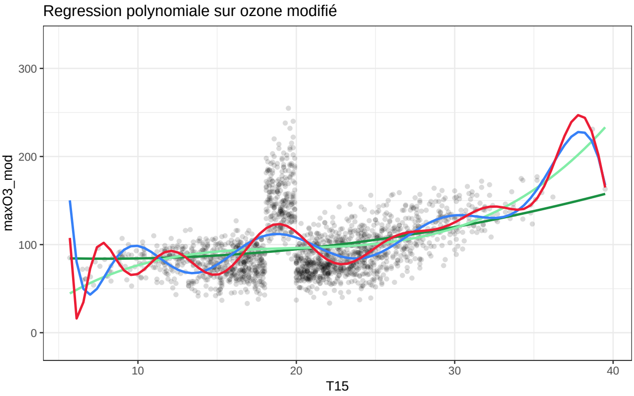
$$Y_i = f(x_i) + E_i$$

$$f(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_d x_i^d$$



Approche paramétrique: régression linéaire polynomiale

Ne prend pas en compte les variations locales



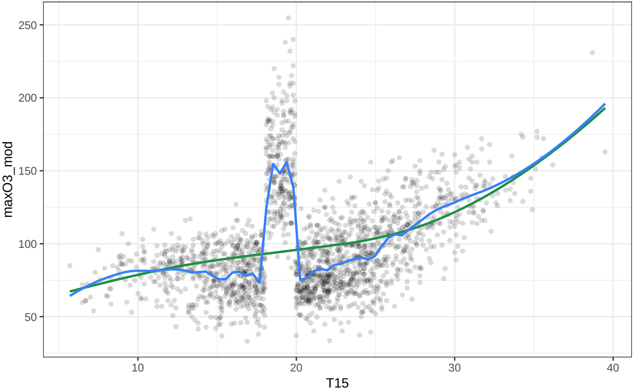
Approche non paramétrique

f n'a plus une forme imposée "Let the data show the appropriate functional form" (Hastie)

Avantage: flexibilité, capte des variations inattendues

Approche non paramétrique

```
p_mod + stat_smooth(method = "loess", se = FALSE, span = 1, col = col_1) +
    stat_smooth(method = "loess", span = 0.1, col = col_3, se = FALSE)
```



Peut s'ajuster très finement aux données : lisseur

Approche non paramétrique

Définition d'un lisseur (Hastie):

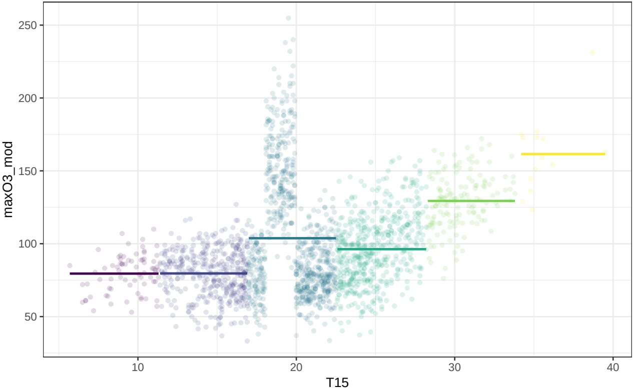
A smoother is a tool for summarizing the trend of a response measurement Y *of one predictor* X_1*. It produces an estimate of the trend that is less variable than* Y *itself; hence the name of smoother.*

- Objectif descriptif
- Estimation de la fonction de régression

Outline

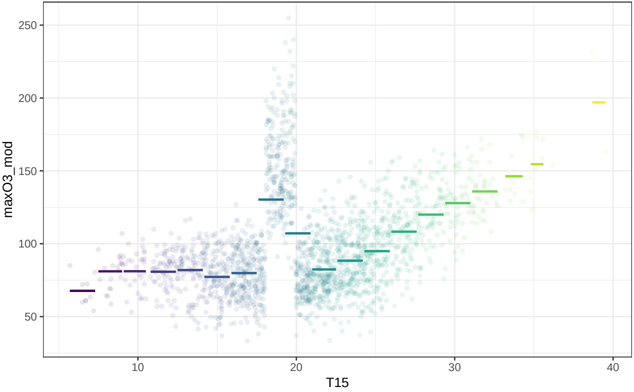
Régressogramme (Bin smoother)

- Découper les x en intervalles réguliers
- Calculer la moyenne des Y dans chaque intervalle



Régressogramme (Bin smoother)

- Découper les x en intervalles réguliers
- Calculer la moyenne des Y dans chaque intervalle



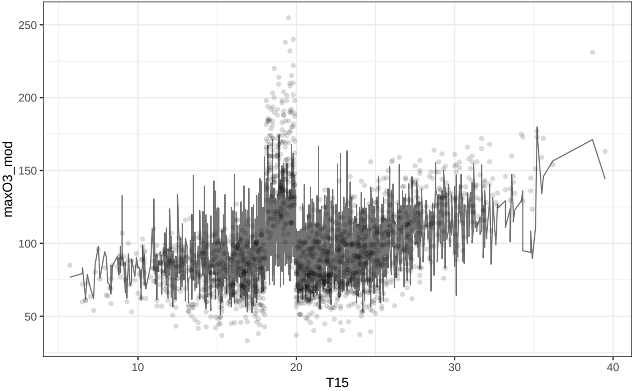
Régressogramme (Bin smoother)

Questions - Choix de la fenêtre (dualité biais - variance) - Problème de discontinuité $\Rightarrow$ Prendre des régions qui se chevauchent

Moyennes mobiles

- Principe: définir, en chaque point, un voisinage pour calculer la moyenne de Y (moyenne sur des intervalles glissants)
- Avantage: simple et intuitif

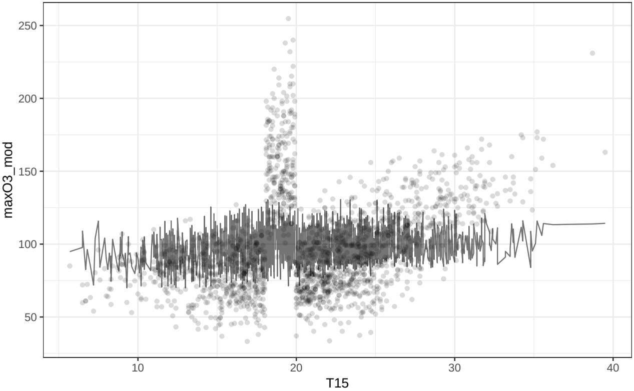
```
p_mod +
  geom_line(aes(y=rollmean(maxO3_mod, k = 5, na.pad=TRUE), alpha=0.9)) + theme(legend.position=
```



Moyennes mobiles

- Principe: définir, en chaque point, un voisinage pour calculer la moyenne de Y (moyenne sur des intervalles glissants)
- Avantage: simple et intuitif

```
p_mod +
  geom_line(aes(y=rollmean(maxO3_mod, k = 30, na.pad=TRUE), alpha=0.9)) + theme(legend.position
```



- Taille du voisinage (dualité biais - variance) - Poids identiques à tous les points

Moyenne mobile pondérée : Nadaraya-Watson

Moyenne mobile calculée par:

$$\frac{\sum_i p(x_i)Y_i}{\sum_i p(x_i)}$$

avec les poids

$$p(x_i) = K\left(\frac{x_i - x_0}{\lambda}\right)$$

Fonction des poids décroissante en $|x - x_0|$ et symétrique - λ : largeur de la fenêtre - λ élevé $\Rightarrow$ les x_i ont le même poids $\Rightarrow$ approximation est lisse - Exemple du noyau gaussien

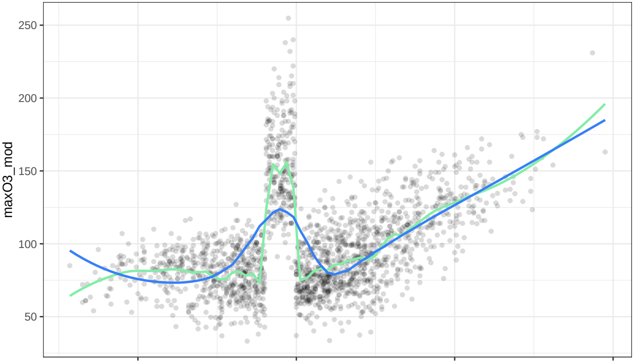
$$p(x_i) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{(x_i - x_0)^2}{2}\right)$$

Régression polynomiale locale pondérée (loess)

Pourquoi se contenter de la moyenne? \ On en veut toujours plus : régression polynomiale locale pondérée

- méthode `loess`
- souvent on se contente de polynôme de degré 2
- choix d'un voisinage autour de x_0 ou plus proches voisins
- `span` : proportion de points constituant le voisinage

```
p_mod  +  stat_smooth(method="loess", span = 0.1, col = col_2, se = FALSE ) + stat_smooth(metho
```



Régression polynomiale locale pondérée (loess)

Paramètre de la méthode

- rayon du voisinage ou proportion (`span`) des points pris en compte dans le lissage -`span` proche de 0 ⇒ interpolation : biais faible, variance forte
- `span` proche de 1 ⇒ régression constante: biais fort, variance faible

Arbitrage entre biais et variance

Choix de fenêtre

Estimation de la fenêtre optimale par apprentissage - validation : - Séparer le jeu de données en proportion 2/3 pour apprentissage et 1/3 pour validation - Faire varier la taille de la fenêtre - Estimer le modèle sur les données d'apprentissage - Calculer la somme des erreurs au carré sur les données de validation - Choisir la fenêtre qui minimise les erreurs de prédiction

Si peu de données $\Rightarrow$ validation croisée

Choix de fenêtre

```r
set.seed(19)
n <- nrow(ozone_mod)
span = 0.1
ind_test <- sample(1:n,size = round(n /10), replace =FALSE )
ozone_train <- ozone_mod %>% filter( ! (row_number()%in% ind_test))
ozone_test <- ozone_mod %>% filter(  (row_number()%in% ind_test))
ozone_loess <- loess(maxO3_mod~T15, data = ozone_train)

error_fit <- sum(ozone_loess$residuals^2)/nrow(ozone_train)

ozone_test <- ozone_test %>% mutate(pred = predict(ozone_loess, newdata = ozone_test),
                     res = maxO3_mod - pred)
error_pred <- sum(ozone_test$res^2)/ nrow(ozone_test)

cat("Erreur d'ajustement : ", error_fit, "\nErreur de prediction : ", error_pred, "\n")
```

```
## Erreur d'ajustement :  815.041
## Erreur de prediction :  938.6973
```

Erreur d'ajustement - erreur de prédiction

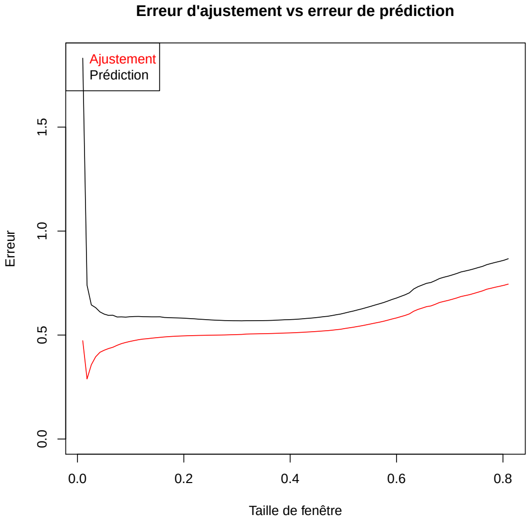


Figure 2:

Splines

Régression polynomiale par morceaux : - nécessité de déterminer les nœuds (les points de jonction): nombre et positions - degré du polynôme (souvent polynôme cubique)

Splines

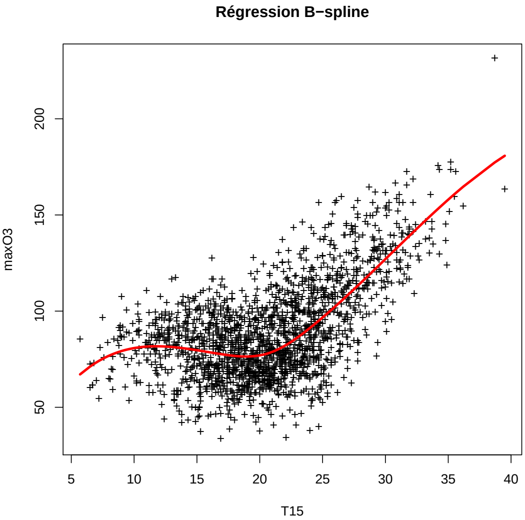


Figure 3:

Outline

Cas multidimensionnel

Modèle paramétrique :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i1} x_{i2} + ... + \beta_j x_{ip} + E_i$$

Extension naturelle au modèle non paramétrique :

$$Y_i = f(x_{i1}, x_{i2}, ..., x_{ip}) + E_i$$

Fléau de la dimension (peu de données dans un voisinage multidimensionnel) $\Rightarrow$ Estimation trop difficile de f

Simplification avec modèle additif :

$$Y_i = f_1(x_{i1}) + f_2(x_{i2}) + ... + f_p(x_{ip}) + E_i$$

Modèle additif

Quel est l'effet d'un facteur sur Y, les autres étant constants?

```
library(gam)
res.gam <- gam(maxO3~lo(T15)+lo(maxO3v),data=ozone)
#plot(res.gam,ask = TRUE)
summary(res.gam)
```

```
##
## Call: gam(formula = maxO3 ~ lo(T15) + lo(maxO3v), data = ozone)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -63.0125  -9.0534   0.2811   9.5980  54.3707
##
## (Dispersion Parameter for gaussian family taken to be 221.1302)
##
##     Null Deviance: 1123344 on 1885 degrees of freedom
## Residual Deviance: 415221.4 on 1877.724 degrees of freedom
## AIC: 15544.54
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                Df Sum Sq Mean Sq F value    Pr(>F)
## lo(T15)       1.0 350271  350271 1584.00 < 2.2e-16 ***
## lo(maxO3v)    1.0 201473  201473  911.11 < 2.2e-16 ***
## Residuals  1877.7 415221     221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
```

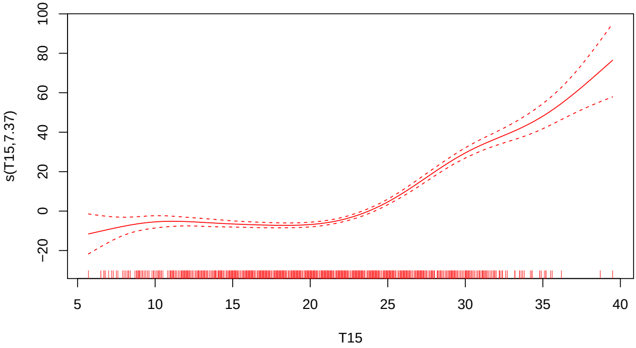
Package mgcv

Package très complet

Utilise principalement les splines

Propose une solution pour le choix délicat des paramètres de lissage par validation croisée généralisée

```
library(mgcv)
res.mgcv = gam(maxO3~s(T15)+s(maxO3v),data=ozone)
plot(res.mgcv,col="red")
```



Choix de modèle

Besoin de sélectionner des variables - Test de modèles emboîtés - Critère AIC ou BIC - Par validation croisée: trouver le modèle à une variable qui prédit le mieux, puis à 2 variables, . . .

```
res.gam <- gam::gam(maxO3~lo(T15)+lo(maxO3v),data=ozone)
res.gam1 <-  gam::gam(maxO3~lo(maxO3v),data=ozone)
anova(res.gam1,res.gam)
```

```
## Analysis of Deviance Table
##
## Model 1: maxO3 ~ lo(maxO3v)
## Model 2: maxO3 ~ lo(T15) + lo(maxO3v)
##   Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)
## 1    1881.1     605476
## 2    1877.7     415221 3.4195   190255 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Outline

Conclusion

Peu de données $\Rightarrow$ faire des hypothèses sur les liaisons (modèles paramétriques)

Beaucoup de données $\Rightarrow$ possibilité d'utiliser des modèles additifs

Problèmes : éviter le surajustement (sélection de variables), choix des paramètres de lissage

Extension aux modèles additifs généralisés (GAM): l'erreur peut ne pas être normale, Y peut être qualitative