Polycopié de travaux dirigés

ENSIIE

Régression avancée et R

chrisotphe.ambroiset@genopole.cnrs.fr

Semestre de printemps 2012
Université d'Évry Val d'Essonne

Table des matières

Première partie

Travaux dirigés

Régression linéaire multiple

Exercice 1.1. Le jeu de données `gavote` décrit le vote présidentiel aux états-unis en 2000, dans l'état de Géorgie. Chacun des 159 "canton" est décrit par les variables suivantes :

- `equip` Le système physique de vote
  - LEVER : machine à levier
  - OS-CC : Scan optique comptage centralisé ("central count"),
  - OS-PC : Scan optique comptage local ("precinct count")
  - PAPER : vote par bulletin papier
  - PUNCH : vote par poinçon
- `econ` le statut économique du "canton" (`middle`, `poor`, `rich`).
- `perAA`, le pourcentage d'afro-américains
- `rural` indicateur de la ruralité du canton (urban, rural)
- `atlanta` indicateur de l'appartenance ou non à Atlanta
- `gore` nombre de votes pour Gore
- `bush` nombre de votes pour Bush
- `other` number of votes for other candidates
- `votes` nombre de votes validés
- `ballots` nombre de bulletins

1. Charger le jeu de données `gavote` et faites un résumé numérique des données.
2. Créer la variable `undercount` qui est la proportion de bulletins de vote considérés comme nuls. Représenter la distribution de cette nouvelle variable. Créer la variable `pergore` (pourcentage de votants pour `Gore`) et tracer le diagramme de dispersion croisant `pergore` avec `perAA`.
3. Tracer la droite de régresssion.
4. Représenter la distribution de la variable `equip`
5. Régresser `undercount` sur `perAA`. Interprétez ce modèle `model1`.
6. Calculer la somme des résidus au carré, le R^2 et le R^2 ajusté.
7. Régresser `undercount` sur `rural`. Interprétez ce modèle `model2`.
8. Centrer les variables `pergore` et `perAA`, et ajuster un modèle (`model3`) de régression linéaire qui explique `undercount` en fonction de `cperAA`, `cpergore`, `rural` et `equip`. Commentez.
9. Supprimer toutes les variables non significatives et comparer le modèle obtenu avec le précédent. Commentez.
10. En utilisant la procédure `step`, simplifier le modèle `model3`. Commentez.
11. Faites le diagnostique de la régression finale.

Travaux dirigés 2

Régression logistique

Exercice 2.1. Considérons le jeu de données `esoph` sur le cancer de l' œsophage en Ile-et-Vilaine L'objectif est de mener une étude cas-témoins [1] et de comprendre quels sont les facteurs influant sur le déclenchement de la maladie.

1. En regardant la structure du tableau de données expliquer la nature de chaque variable.
2. Faites un résumé numérique du tableau.
3. Créer un tableau de contingence à 2 lignes, 4 colonnes dont chaque case contient le nombre de cancer et les contrôles en fonction de leur consommation de tabac. Dessinez un diagramme mosaïque [2] du tableau.
4. Estimer un modèle `model0` de régression logistique en utilisant la seule variable `tobgp` (consommation de tabac).
5. Interpréter les coefficients en terme d'odd ratios (rapport de chance).
6. En recodant la première modalité comme non fumeur et fumeur pour les autres, estimer un modèle `model1` et calculer l'effet de la cigarette.
7. Estimer un modèle `model2` de régression logistique en utilisant la seule variable `tobgp` (consommation d'alcool).
8. Estimer un modèle `model3` de régression logistique en utilisant toutes les variables et interactions.
9. Utiliser la fonction `unclass` pour transformer toutes vos variables qualitatives en variables quantitatives. Estimer un dernier modèle `model4` avec ces dernières variables
10. Considérons le model `model4` : quel est la probabilité q'un homme de 25 ans qui ne boit pas ni ne fume développe un cancer de l'œsophage.
11. Faites une analyse graphique des résidus

1. Un groupe de personnes atteintes d'une maladie (cas) est comparé à un groupe de sujets qui n'ont pas la maladie étudiée (témoins). Le but est la recherche d'un ou des facteurs d'exposition antérieurs à la maladie susceptibles de pouvoir l'expliquer. Ce type d'étude sert donc à tester une hypothèse spécifique avec une association d'un facteur de risque.

2. Ce diagramme est composé d'autant de rectangles (et/ou carrés) qu'il y a cellules dans le tableau de contingence de départ. Pour la représentation de chaque mosaïque la largeur de la bande sera proportionnelle aux fréquences marginales. La hauteur de chaque mosaïque sera proportionnelle au rapport de l'effectif de la cellule sur le total de la colonne.

Travaux dirigés 3

Régression non paramètrique et modèle additif

Exercice 3.1. Considèrons le jeu de données `uswages` issu d'une étude de 1988. Regardons les variables nombre d'années d'éducation (`educ`) et salaire (`wage`).

1. Calculez un premier estimateur du salaire en fonction du nombre d'années d'études :
2. Calculez un second estimateur du salaire en fonction du nombre d'année en utilisant un estimateur à noyau (via `ksmooth`).
3. Comparer l'erreur quadratique moyenne sur les données d'apprentissage des deux estimations précédentes et commentez.
4. Calculez l'erreur quadratique moyenne en validation croisée. Commentez.

Exercice 3.2. Considérons les données `kyphosis`, qui décrivent les résultat d'une chirurgie corrective sur 81 enfants atteints d'une déformation de la colonne vertébrale. Ajuster un modèle additif et pour prédire la variable réponse `Kyphosis`, qui vaut un lorsqu'une déformation est présente après l'opération et zéro sinon. Commentez.

Projet

Le jeu de données `dvisists` provient d'un étude du ministère de la santé Australien réalisée en 1977-1978. Pour accéder aux données, il vous suffit d'installer le module `faraway` et charger ensuite les données :

```
library(faraway)
data(dvisits)
```

1. Expliquer en quelques lignes la nature d'une régression log-linéaire ainsi que son intérêt.
2. Construire un modèle de regression de Poisson (log-linéaire) avec pour variable réponse le dvisists et comme variables explicatives `sex`, `age`, `agesq`, `income`, `levyplus`, `freepor`, `freerepa`, `illness`, `actdays`, `hscore`, `chcond1`, `chcond2`.

   Est-ce que ce modèle est un modèle raisonnable (utiliser la déviance pour répondre à cette question).
3. Afficher les résidus versus les données estimées. Comment expliquez vous les lignes sur le graphique affiché ?
4. Utiliser une procédure d'élimination descendante avec un seuil à 5% pour réduire la taille de votre modèle autant que possible.
5. Quel type de personne serait d'après votre modèle la plus susceptible de consulter le médecin ?
6. Pour la dernière personne du jeu de données prédire la probabilité de visite chez le docteur 0, 1, ou 2 fois.
7. Utiliser un modèle gaussien pour résoudre le problème. Décrire les différences.
8. Proposer et tester des modèles / méhodes de régression aternatives (non paramètrique par exemple) pour améliorer votre modèle.
9. Donner une estimation de l'erreur de prédiction de vos modèles.

Deuxième partie

Corrections

Régression logistique

Solution de l'exercice 2.1. Chargeons les données et calculons le résumé numérique :

```
> library(faraway)
> data(esoph)
> require(graphics) # for mosaicplot
> summary(esoph)
```

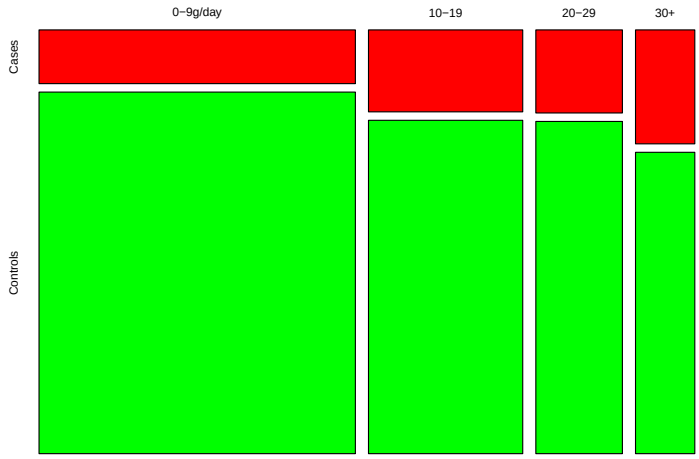
```
   agegp          alcgp         tobgp        ncases         ncontrols
 25-34:15   0-39g/day:23   0-9g/day:24   Min.   : 0.000   Min.   : 1.00
 35-44:15   40-79    :23   10-19   :24   1st Qu.: 0.000   1st Qu.: 3.00
 45-54:16   80-119   :21   20-29   :20   Median : 1.000   Median : 6.00
 55-64:16   120+     :21   30+     :20   Mean   : 2.273   Mean   :11.08
 65-74:15                                3rd Qu.: 4.000   3rd Qu.:14.00
 75+  :11                                Max.   :17.000   Max.   :60.00
```

Le jeu de données comprend 88 entrées qui comprenne chacune le nombre de cas (cancer) et témoins pour des combinaisons des variables age/consommation d'alcool/consommation de tabac.

Le résumé numérique nous permet de voir l'ensemble des modalités de chacune des variables, ainsi que des statistiques sur le nombre de cas et de témoins dans chacune des catégories ainsi définies.

La construction du tableau de contingence à 2 lignes, 4 colonnes dont chaque case contient le nombre de cancer et les contrôles en fonction de leur consommation de tabac peut être réalisé en faisant une boucle mais peut être traitée relativement rapidement en R.

```
> tobtab<-rbind(tapply(esoph$ncases,esoph$tobgp,sum),tapply(esoph$ncontrols,esoph$tobgp,sum))
> rownames(tobtab)<-c("Cases","Controls")
> plot(as.table(t(tobtab)),col=c("red","green"),main="")
```



Le modèle de régression logistique en utilisant la seule variable `tobgp` (consommation de tabac) est estimé comme suit :

```
> model0 <- glm(cbind(ncases, ncontrols) ~  tobgp, data = esoph,
+                 family = binomial(),
+                 contrasts=list(tobgp='contr.treatment'))
> summary(model0)

Call:
glm(formula = cbind(ncases, ncontrols) ~ tobgp, family = binomial(),
    data = esoph, contrasts = list(tobgp = "contr.treatment"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0770  -1.2018   0.2108   1.0230   3.7459

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9067     0.1213 -15.713  < 2e-16 ***
tobgp10-19    0.5033     0.1903   2.645 0.008166 **
tobgp20-29    0.5204     0.2294   2.269 0.023272 *
tobgp30+      0.9340     0.2433   3.839 0.000123 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 227.24  on 87  degrees of freedom
Residual deviance: 209.53  on 84  degrees of freedom
AIC: 365.01

Number of Fisher Scoring iterations: 5
```

Les contrastes sont des recodages de variables qualitatives. Chaque type de contraste s'interprète de manière spécifique. Le codage interne peut être accéder un regardant de plus près la matrice de design X.

```
> head(model.matrix(model0))

  (Intercept) tobgp10-19 tobgp20-29 tobgp30+
1           1          0          0        0
```

```
2          1          1          0          0
3          1          0          1          0
4          1          0          0          1
5          1          0          0          0
6          1          1          0          0
```

Dans notre cas la classe de référence est la classe `tobgp0-9g` qui est codée (000).

Dans notre modèle L'intercept β_0 sera donc interprété comme un rapport des chance moyen entre probabilité d'avoir un cancer et ne pas avoir un cancer quelque que soit la catégorie de fumeur à laquelle on appartient.

Le -1 est la convention qui permet de supprimer l'estimation de 'intercept.

On peut voir qu'un coefficient est associé à chaque modalité :

La variable `tobgp`, qui possède quatre modalités est code par 3 coordonnées (les contrastes de traitement).

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 \mathbb{I}_{x_i = tobgp10-19} + \beta_2 \mathbb{I}_{x_i = tobgp20-29} + \beta_3 \mathbb{I}_{x_i = tobgp30+}$$

Pour tester la validité du modèle, on réalise un test du rapport de vraisemblance, qui teste la déviance :

$$\begin{cases} H_0 : \text{le modèle vide est correct} \\ H_1 : \text{le modèle proposé est meilleur} \end{cases}$$

Le modèle vide suppose que la probabilité d'avoir un cancer ne dépend pas de la consommation de tabac. Sous H_0 l'écart de déviance entre le modèle vide et le modèle proposé suit un χ^2 à $s-1$ degrés de liberté (s étant le nombre de paramètres du modèle proposé et 1 et le nombre de paramètre du modèle vide). La pvalue est la probabilité d'observer un plus grand ecart de déviance que celui constatée, sous H_0. Si cette pvalue est supérieure au seuil fixé alors on accepte H_0 s

```
> model_vide <- glm(cbind(ncases, ncontrols) ~ 1, data = esoph, family = binomial)
> anova(model_vide, model0, test="Chisq")

Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ 1
Model 2: cbind(ncases, ncontrols) ~ tobgp
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        87     227.24
2        84     209.53  3   17.709 0.0005049 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On décide donc H_1 : le modèle a un intérêt au global..

Ensuite en regardant les résultats des tests de Wald pour chaque coefficient, on remarque que les quatre coefficients semblent significatifs.

L'exponentiel des paramètres peut être interprété en terme de rapport des chance (rapport sur la probabilité d'appartenir à la classe par rapport à celle de ne pas appartenir à cette classe).

```
> exp(coef(model0))

(Intercept)  tobgp10-19  tobgp20-29    tobgp30+
  0.1485714   1.6541721   1.6826923   2.5445591
```

On voit ce que montrait déjà le diagramme mosaïque. Plus l'on fume et plus la probabilité d'avoir un cancer de l'œsophage est importante.

En recodant la première classe comme non fumeur contre fumeur pour les autres, on peut calculer l'effet de la cigarette au global :

```
> esoph$smokegp<-factor(c("NotSmoker",rep("Smoker",3))[as.numeric(esoph$tobgp)])
> model1 <- glm(cbind(ncases, ncontrols) ~  smokegp,
+                           data = esoph, family = binomial())
> summary(model1)

Call:
glm(formula = cbind(ncases, ncontrols) ~ smokegp, family = binomial(),
    data = esoph)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0770  -1.3268   0.1252   0.9432   3.7459

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.9067     0.1213 -15.713  < 2e-16 ***
smokegpSmoker   0.6015     0.1586   3.793 0.000149 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 227.24  on 87  degrees of freedom
Residual deviance: 212.52  on 86  degrees of freedom
AIC: 364

Number of Fisher Scoring iterations: 5

> exp(coef(model1))

  (Intercept) smokegpSmoker
    0.1485714     1.8247863
```

et l'on s'apercoit que fumer est associé à une probabilité 1.8 fois plus grande de cancer de l'œsophage.

En identifiant le `model2` de régression logistique en utilisant la seule variable `tobgp` (consommation d'alcool),

```
>  model2 <- glm(cbind(ncases, ncontrols) ~  alcgp,
+                           data = esoph, family = binomial(),
+                          contrasts=list(alcgp='contr.treatment'))
> summary(model2)

Call:
glm(formula = cbind(ncases, ncontrols) ~ alcgp, family = binomial(),
    data = esoph, contrasts = list(alcgp = "contr.treatment"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6629  -1.0478  -0.0081   0.6307   3.0296

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6610     0.1921 -13.854  < 2e-16 ***
alcgp40-79    1.1064     0.2303   4.804 1.56e-06 ***
alcgp80-119   1.6656     0.2525   6.597 4.20e-11 ***
alcgp120+     2.2630     0.2721   8.317  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 227.24  on 87  degrees of freedom
Residual deviance: 138.79  on 84  degrees of freedom
AIC: 294.27

Number of Fisher Scoring iterations: 5
```

on s'apercoit que le modèle semble légèrement plus intéressant que le précédent

```
> ## changement Guillem
> #pchisq(deviance(model2),df.residual(model2),lower=FALSE)
> anova(model0, model2, test="Chisq")

Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ tobgp
Model 2: cbind(ncases, ncontrols) ~ alcgp
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        84     209.53
2        84     138.79  0   70.742
```

est meilleur que le `model0` (si l'on considère les pvalues) et les critères AIC :

```
> print(AIC(model2))

[1] 294.27

> print(AIC(model0))

[1] 365.0118
```

En considérant l'ensemble des variables nous pouvons construire un diagramme mosaïque cas-témoins pour toute les combinaisons de modalités disponibles :

```
>  ttt <- table(esoph$agegp, esoph$alcgp, esoph$tobgp)
>  ttt[ttt == 1] <- esoph$ncases
>  tt1 <- table(esoph$agegp, esoph$alcgp, esoph$tobgp)
>  tt1[tt1 == 1] <- esoph$ncontrols
>  tt <- array(c(ttt, tt1), c(dim(ttt),2),
+              c(dimnames(ttt), list(c("Cancer", "control"))))
> mosaicplot(tt, main = "esoph data set", color=c("red","green")
+ )
```

En faisant un modèle qui prend en compte tous les effet et interaction on obtient le modèle `model3` :

```
> model3 <- glm(cbind(ncases, ncontrols) ~ agegp +tobgp * alcgp,
+                          data = esoph, family = binomial())
```

Il est effectivement beaucoup plus complexe :

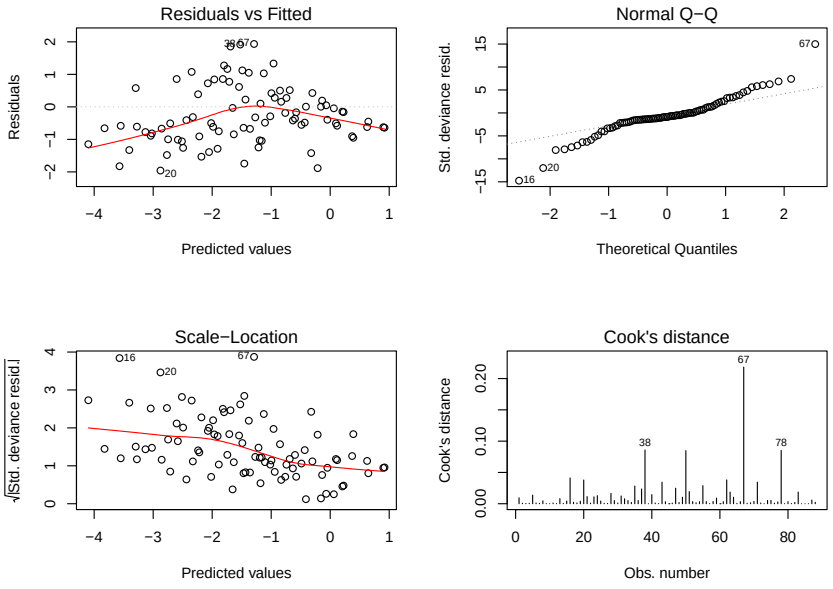
```
> ### changement Guillem
> ###pchisq(deviance(model3),df.residual(model3),lower=FALSE)
> anova(model0, model2, model3, test="Chisq")

Analysis of Deviance Table

Model 1: cbind(ncases, ncontrols) ~ tobgp
Model 2: cbind(ncases, ncontrols) ~ alcgp
Model 3: cbind(ncases, ncontrols) ~ agegp + tobgp * alcgp
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        84    209.531
2        84    138.789  0   70.742
3        67     47.484 17   91.305 3.531e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>  AIC(model3)

[1] 236.9645
```

En utilisant la routine `stepAIC` de la library `MASS`, les modèles explorés à partir du `model3` sont toujours moins bon que le `model0`.

```
> model4 <- glm(cbind(ncases, ncontrols) ~ unclass(agegp) + unclass(tobgp) + unclass(alcgp), data = esoph, family = bi
```

Pour prédire la probabilité q'un homme de 25 ans qui ne boit pas ni ne fume développe un cancer de l'œsophage, nous pouvons fabriquer un tel individu et utiliser la régression logistique :

```
> logitpi<-predict(model4,newdata=list(agegp=1,
+                              alcgp = 1,
+                              tobgp=1 ))
> exp(logitpi)/(1+exp(logitpi))

         1
0.01631861

> ilogit(logitpi)

         1
0.01631861


> par(mfrow=c(2,2))
> plot(model4, which = 1:4)
```



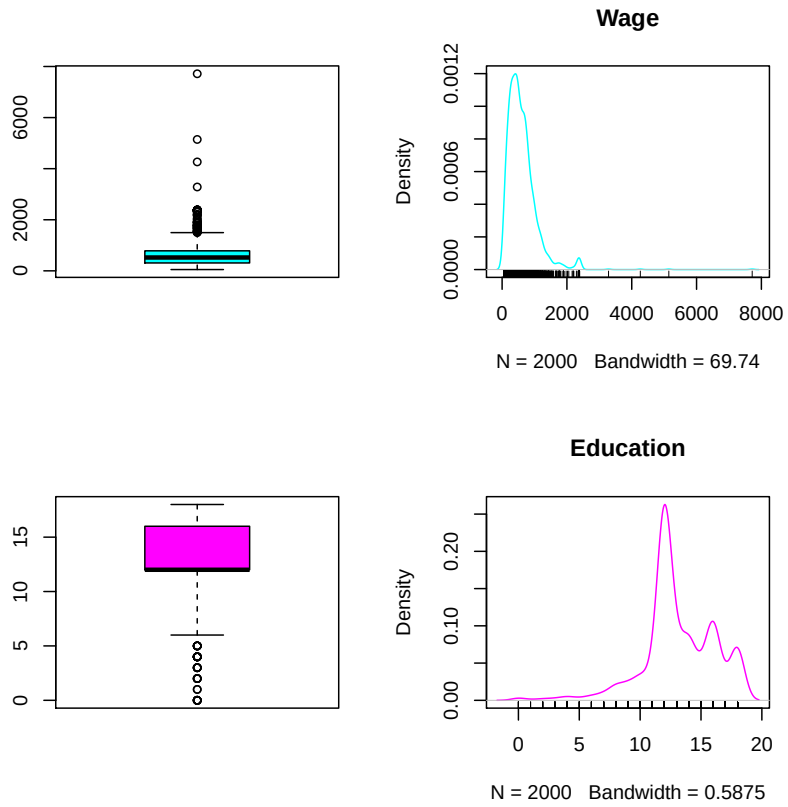
Éléments de corrections 3

Régression non paramètrique et modèle additif

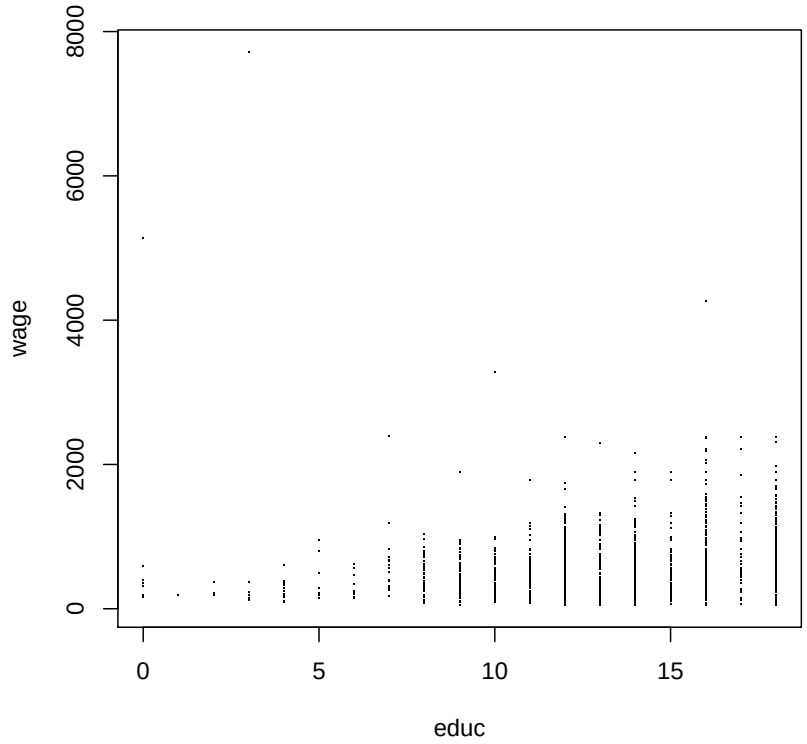
Solution de l'exercice 3.1. Considèrons le jeu de données `uswages` issu d'une étude de 1988.

Regardons les variables nombre d'années d'éducation (`educ`) et salaire (`wage`)

```
> data(uswages)
> summary(uswages$wage)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  50.39  308.60  522.30  608.10  783.50 7716.00

> summary(uswages$educ)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   12.00   12.00   13.11   16.00   18.00

> par(mfrow = c(2, 2))
> boxplot(uswages$wage, col = 5)
> plot(density(uswages$wage), col = 5, main = "Wage")
> rug(uswages$wage)
> boxplot(uswages$educ, col = 6)
> plot(density(uswages$educ), col = 6, main = "Education")
> rug(uswages$educ)
```

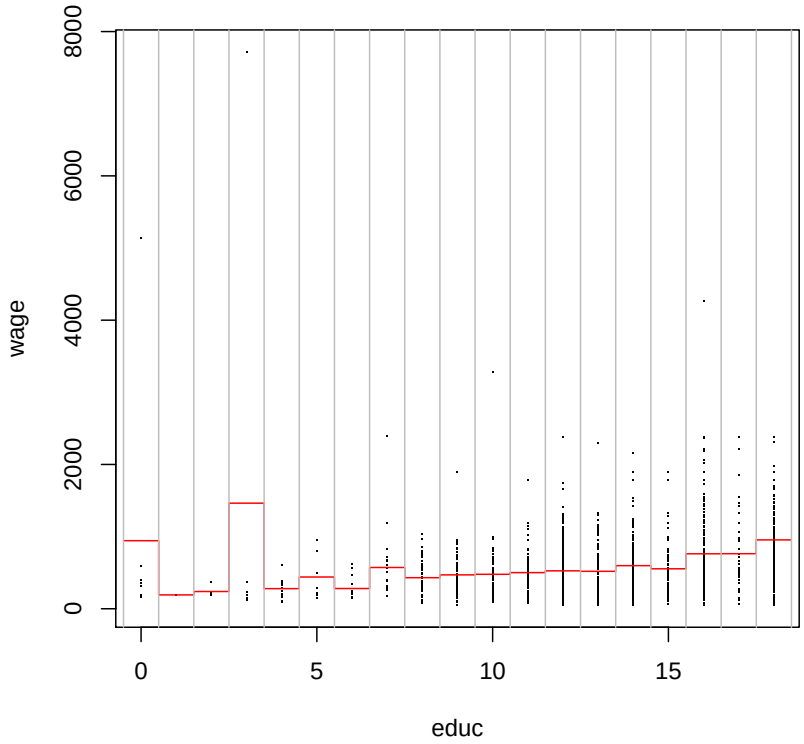



```
> plot(wage ~ educ, data = uswages, pch = ".")
```



Calculons un premier estimateur du salaire en fonction du nombre d'années d'études :

```
> plot(wage ~ educ, data = uswages, pch = ".")
> model1 <- tapply(uswages$wage, uswages$educ, mean)
> lines(c(-0.5, rep(0.5:17.5, each = 2), 18.5), rep(model1, each = 2),
+     type = "l", col = 2)
> abline(v = -0.5:18.5, col = "gray")
```

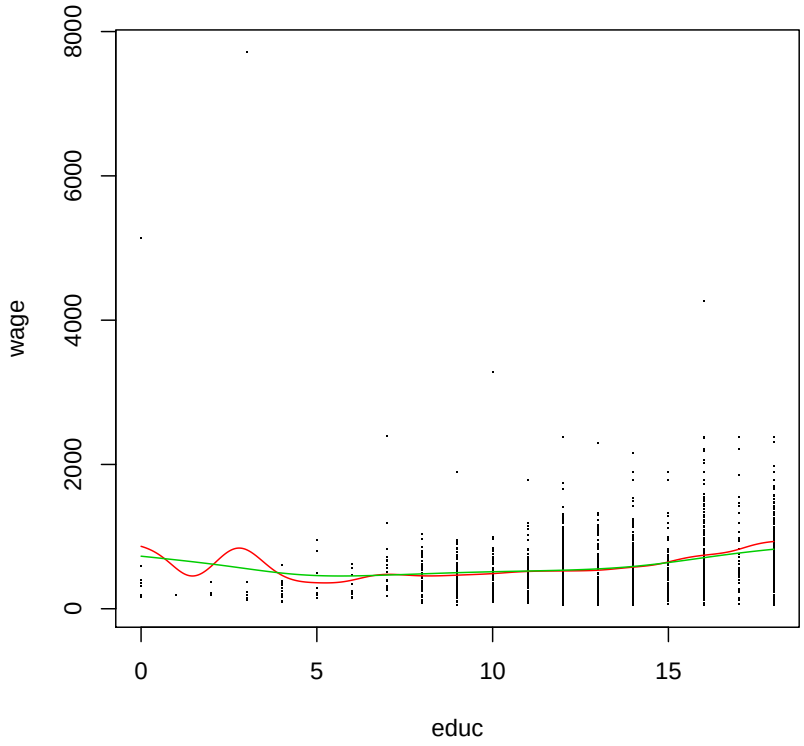



```
> attach(uswages)

The following object(s) are masked from 'uswages (position 4)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 5)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 6)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 8)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 9)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 10)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 11)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 12)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 13)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 14)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 15)':
```

```
    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 16)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 17)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 18)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 19)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 20)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we
The following object(s) are masked from 'uswages (position 21)':

    educ, exper, mw, ne, pt, race, smsa, so, wage, we

> n <- length(wage)
> predict.mean <- function(x, model1) {
+     model1[x + 1]
+ }
> MSE.model1 <- 1/n * sum((predict.mean(educ, model1) - wage)^2)
```

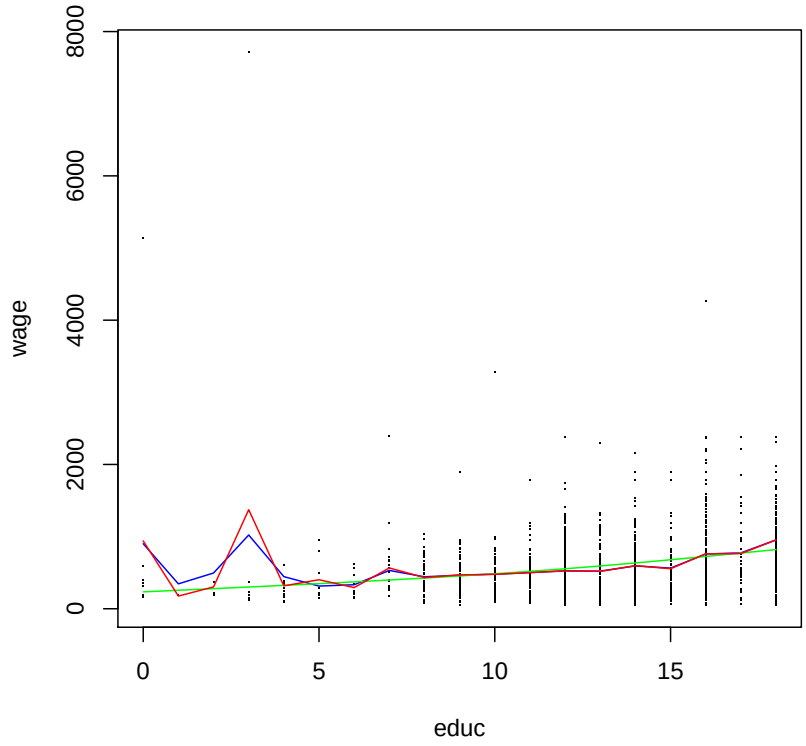
Calculons un second estimateur :

```
> plot(wage ~ educ, pch = ".")
> lines(ksmooth(educ, wage, "normal", bandwidth = 2), col = 2)
> lines(ksmooth(educ, wage, "normal", bandwidth = 5), col = 3)
> MSE.model2 <- 1/n * sum((ksmooth(educ, wage, "normal", bandwidth = 2,
+     x.points = educ)$y - wage[order(educ)])^2)
> print((MSE.model1 - MSE.model2)/MSE.model2)

[1] -0.01490381
```



On peut constater que l'estimation obtenue par la méthode de la moyenne semble meilleure en terme de MSE.

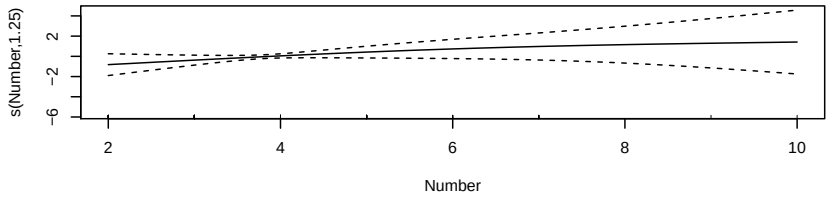
```
> plot(wage ~ educ, pch = ".")
> lines(smooth.spline(educ, wage, spar = 0.1), col = "blue")
> lines(smooth.spline(educ, wage, spar = 0.8), col = "green")
> lines(model3 <- smooth.spline(educ, wage), col = "red")
> print(MSE.model3 <- 1/n * sum((predict(model3, educ)$y - wage)^2))

[1] 185579.4
```

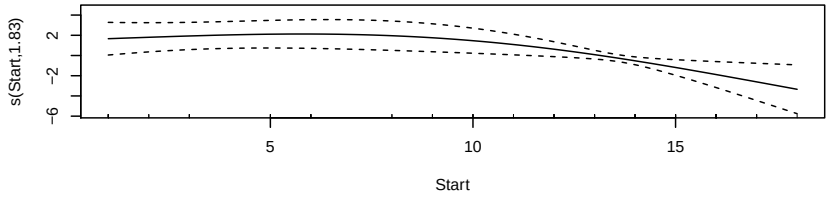


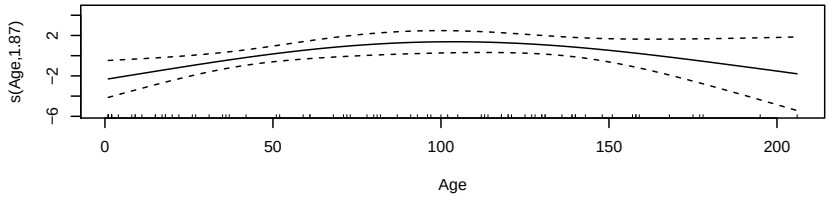
Calculons maintenant ces même deux erreur en validation croisée.

Solution de l'exercice 3.2. Ajuster un modèle additif et pour prédire la variable réponse `Kyphosis` et commentez.

```
> library(mgcv)
> model.gam <- gam(Kyphosis ~ s(Number, k = 3) + s(Start, k = 3) +
+     s(Age, k = 3), data = kyphosis, family = binomial())
> summary(model.gam)

Family: binomial
Link function: logit

Formula:
Kyphosis ~ s(Number, k = 3) + s(Start, k = 3) + s(Age, k = 3)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.5147     0.5714  -4.401 1.08e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
            edf Ref.df Chi.sq p-value
s(Number) 1.251  1.439  1.938  0.2563
s(Start)  1.830  1.971  8.928  0.0111 *
s(Age)    1.869  1.982  6.348  0.0411 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.367   Deviance explained = 41.1%
UBRE score = -0.24784  Scale est. = 1         n = 81

> plot(model.gam, pages = 1)
```


s(Number,1.25)
2
−2
−6
2
4
6
8
10
Number


s(Start,1.83)
2
−2
−6
5
10
15
Start


s(Age,1.87)
2
−2
−6
0
50
100
150
200
Age