TP d'Econométrie II avec le logiciel R

Régression logistique et polytomique

L'objectif de ce TP est de se familiariser avec l'étude de variable qualitative en fonction d'autres variables. On se centre dans ce TP sur le cas de la régression logistique puis sur celui de la régression polytomique.

Régression logistique

On commence par étudier un jeu de données réelles: il s'agit de l'admission en université aux USA suivant les notes obtenues et l'université visée. La variable à expliquer, Y, est notée `admit` dans le jeu de données. Deux notes, `GRE` (Greater Record Exam score) et `GPA`, (Grade Point average) sont des variables pouvant expliquer Y, tout comme `rank`, qui est un indice de 1 à 4 "classant" l'Université visée (4 pour les meilleures universités à 1 pour les moins cotées). On télécharge ce jeu de données et on effectue une régression logistique sur ce jeu. On commence par avoir quelques idées sur les données par le jeu de commandes suivantes:

```
mydata=read.csv("https://www.idre.ucla.edu/stat/data/binary.csv")
attach(mydata)
names(mydata)
summary(gre)
sd(gre)
summary(gpa)
sd(gpa)
table(rank)
table(admit)
table(rank,admit)
```

Expliquer toutes les commandes effectuées et les données affichées.
Ensuite on effectue une régression logistique:

```
logit1=glm(admit~gre+gpa+as.factor(rank),family=binomial(link="logit"),na.action=na.pass)
summary(logit1)
confint(logit1)
```

Quels sont les résultats obtenus? Essayer ensuite en considérant la variable `rank` comme une variable quantitative. Quel modèle préferer?

```
library(aod)
wald.test(b=coef(logit1),Sigma=vcov(logit1),Terms=4:6)
l=cbind(0,0,0,1,-1,0)
wald.test(b=coef(logit1),Sigma=vcov(logit1),L=l)
```

On effectue donc des tests de Wald. Que teste-t-on exactement?
Des tests du rapport de vraisemblance sont aussi possibles:

```
library(lmtest)
lrtest(logit1)
lrtest(logit1,2)
lrtest(logit1,c(2:3))
```

Quelles sont les conclusions? Effectuer les mêmes tests que ceux de Wald et comparer les résultats obtenus lorsque cela est possible.
On peut également effectuer des prédictions à partir du modèle de régression logistique estimé:

```
rank=c(1,2,3,4)
gre=c(mean(mydata$gre))
gpa=c(mean(mydata$gpa))
newdata1=data.frame(gre,gpa,rank)
newdata1
newdata1$rankP=predict(logit1,newdata=newdata1,type="response")
newdata1
newdata2=data.frame(gre=seq(200,800,100),gpa=mean(mydata$gpa),rank=2)
newdata2$greP=predict(logit1,newdata=newdata2,type="response")
cbind(newdata2$gre,newdata2$greP)
```

Quelles sont les résultats des prédictions? Enlever les 100 dernières données de la base de données, réestimer le modèle sur les données restantes et prédire les valeurs de la variable `admit` pour les 100 données retirées. Quelle erreur a-t-on fait? Recommencer la démarche en utilisant un modèle "probit" plutôt que "logit". Quel modèle préférer en regard des prédictions?

Regression polytomique

On commence par reprendre les données précédentes en essayant de modéliser cette fois-ci le rang de l'université visé en fonction des autres variables. On peut ainsi mettre en place une régression polytomique par les commandes suivantes:

```
library(VGAM)
poly_rank=vglm(rank ~gre+gpa+admit,family=multinomial,data=mydata,na.action=na.pass)
summary(poly_rank)
poly_rank@coefficients
poly_rank@fitted.values
poly_rank@y
```

Quels sont les résultats de ces différentes commandes?
On peut également prédire les probabilités d'occurence:

```
admit=as.factor(c(0,1))
gre=c(mean(mydata$gre))
gpa=c(mean(mydata$gpa))
newdataP=data.frame(gre,gpa,admit)
newdataP_Pred=predict(poly_rank,newdata=newdataP,type="response")
newdataP_Pred
```

Mettre en place les commandes pour obtenir un rang prédit.

On peut également utiliser le fait que le rang soit ordonné pour le modéliser. Cela peut se faire par les commandes suivantes:

```
poly_rank2=vglm(rank~gre+gpa+admit,family=propodds,data=mydata,na.action=na.pass)
summary(poly_rank2)
poly_rank2@coefficients
poly_rank2@fitted.values
poly_rank2@y
```

Comparer les résultats avec les précédents.