Feuille de Travaux Dirigés n$^{\text{o}}$ 8
Régression linéaire simple

Exercice VIII.1. Étude de la pollution de l'air. Cet exercice est issu du livre « Statistiques avec `R` », Pierre-André Cornillon et autres, aux éditions PUR, 2008.

1. Récupérer les données dans R en exécutant les instructions suivantes :
`>ozone<-read.table("...ozone.txt",header=T)`
2. Quelles sont les différentes variables ? Quelle est leur nature ? Obtenir les statistiques descriptives du jeu de données.
3. Nous allons nous intéresser plus particulièrement à deux variables du jeu de données : la variable « `maxO3` » et la variable « `T12` ». Pour cela, nous allons calculer les statistiques élémentaires sur ces deux variables, en utilisant la commande `summary`
`>summary(ozone[,c(« maxO3 »,« T12 »)])`
4. Nous allons maintenant représenter le nuage de points (x_i, y_i). Pour cela, exécuter la ligne de commande suivante :
`>plot(maxO3~T12,data=ozone,pch=15,cex=.5)`
5. Nous allons maintenant estimer les paramètres du modèle de régression linéaire. Pour cela, nous allons utiliser la fonction `lm` (lm pour linear model). Cette fonction permet d'ajuster un modèle linéaire. Exécuter les lignes de commande suivantes :
`>reg.simple<-lm(maxO3~T12,data=ozone)`
`>summary(reg.simple)`
6. Nous pouvons consulter la liste des différents résultats de l'objet `reg.simple` avec :
`>names(reg.simple)`
7. Nous pouvons récupérer les coefficients avec :
`>reg.simple$coef`
ou en utilisant la fonction `coef`, ce qui est la méthode recommandée :
`>coef(reg.simple)`
8. **Remarque :** Nous avons auparavant construit un modèle avec constante. Si jamais la constante n'était pas significative ou si encore le modèle ne doit pas contenir de constante, alors nous devons procéder de la manière suivante pour le construire :
`>reg.ss.constante<-lm(maxO3~T12-1,data=ozone) >reg.ss.constante`
9. Nous allons maintenant tracer la droite de régression. Nous pouvons simplement appliquer la commande `abline(reg.simple)` mais nous préférons nous restreindre au domaine d'observation de la variable explicative. Nous créons donc une grille de points sur les abscisses et appliquons le modèle trouvé sur

cette grille :

```
>plot(max03~T12,data=ozone,pch=15,cex=.5)
>grillex<-seq(min(ozone[,« T12 »]),max(ozone[,« T12 »]),
+length=100)
>grilley<-reg.simple$coef[1]+reg.simple$coef[2]*grillex
>lines(grillex,grilley,col=2)
```

10. Nous allons maintenant analyser les résidus. Les résidus sont obtenus par la fonction `residuals`, cependant les résidus obtenus ne sont pas de même variance (hétéroscédastiques). Nous allons utiliser alors les résidus studentisés, qui eux sont de même variance.

```
>res.simple<-rstudent(reg.simple)
>plot(res.simple,pch=15,cex=.5,ylab=« Résidus »,ylim=c(-3,3))
>abline(h=c(-2,0,2),lty=c(2,1,2))
```

11. Pour terminer nous allons donner la procédure pour prévoir une nouvelle valeur. Ayant une nouvelle observation `xnew`, il suffit d'utiliser les estimations pour prévoir la valeur de Y correspondante. Cependant, la valeur prédite est de peu d'intérêt sans l'intervalle de confiance associé. Voyons cela sur un exemple. Nous disposons d'une nouvelle observation de la température `T12` égale à 19 degrés pour le 1er octobre 2001.

```
>xnew=19
>xnew<-as.data.frame(xnew)
>colnames(xnew)<-« T12 »
>predict(reg.simple,xnew,interval=« pred »)
```

Il faut noter que l'argument `xnew` de la fonction `predict` doit être un data-frame avec les mêmes noms de variables explicatives (ici `T12`). La valeur prévueest `76.5` et l'intervalle de prévision à 95% est `[41.5,111.5]`. Pour représenter sur un même graphique l'intervalle de confiance d'une valeur lissée et l'intervalle de confiance d'une prévision, nous calculons ces intervalles pour l'ensemble des points ayant servi à dessiner la droite de régression. Nous faisons figurer les deux sur le même graphique.

```
>grillex.df<-data.frame(grillex)
>dimnames(grillex.df)[[2]]<-« T12 »
>ICdte<-predict(reg.simple,new=grillex.df,interval=« conf »,
+level=0.95)
>ICprev<-predict(reg.simple,new=grillex.df,interval=« pred »,
+level=0.95)
>plot(max03~T12,data=ozone,pch=15,cex=.5)
>matlines(grillex,cbind(ICdte,ICprev[,-1]),lty=c(1,2,2,3,3),
+col=1)
>legend(« topleft »,lty=2 :3,c(« prev »,« conf »))
```