Notes de cours
Économétrie 1

Shuyan LIU
Shuyan.Liu@univ-paris1.fr
http ://samm.univ-paris1.fr/Shuyan-LIU-Enseignement

Année 2014-2015

Chapitre 1

Introduction

Qu'est-ce que l'économétrie ? À quoi sert - elle ?

– validation de la théorie économique

– investigation

a. mise en évidence de relation entre les variables

b. inférence statistique

c. prévision

Ce cours consiste en trois chapitres

a. Régression linéaire simple (données quantitatives)

b. Régression linéaire multiple (données quantitatives)

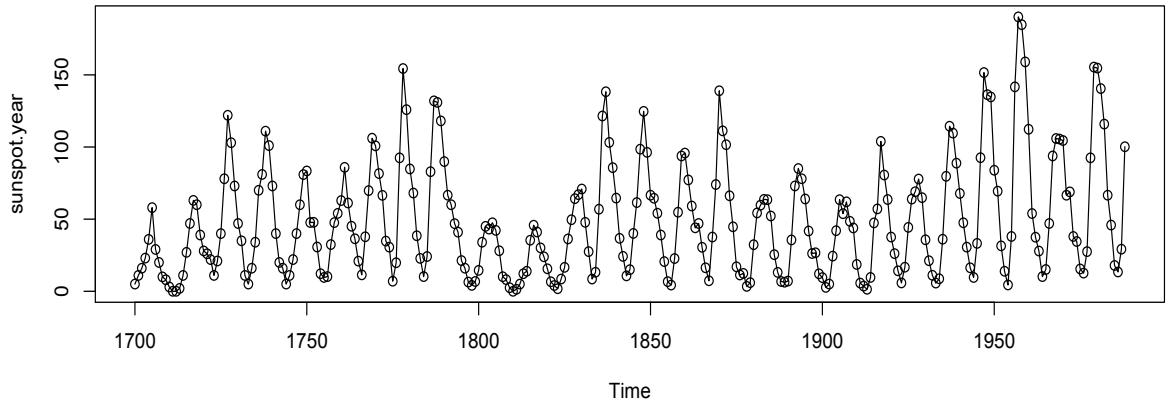
c. Analyse de la variance (données qualitatives)

Quelques exemples

Ces données sont disponibles dans le logiciel R sous les noms : sunspot.year, AirPassengers et chickwts.

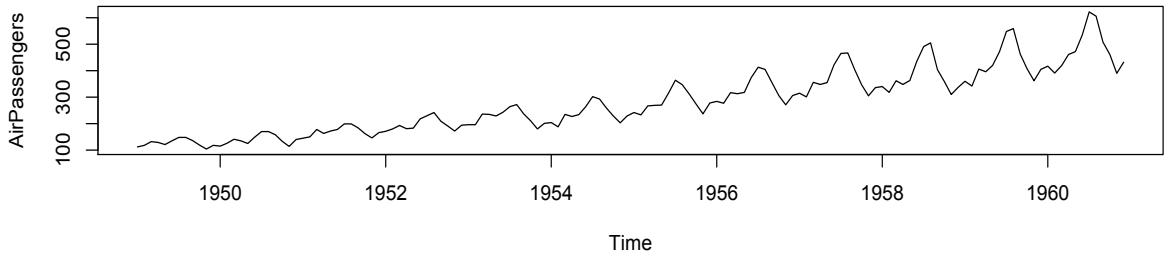
1. Les taches solaires : superposition de deux phénomènes périodiques

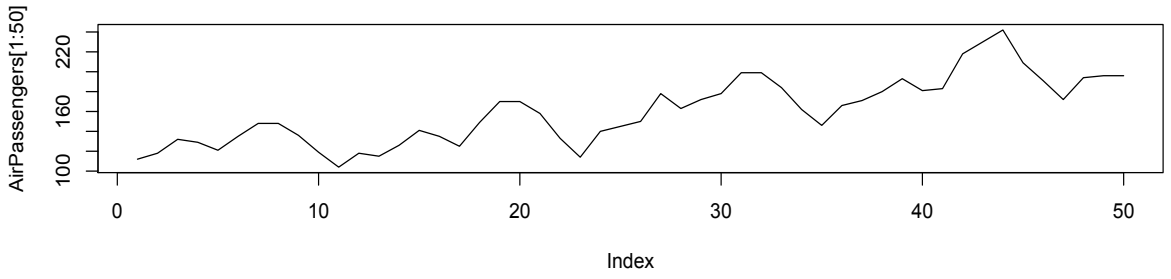
```
> plot(sunspot.year)
> points(sunspot.year)
```



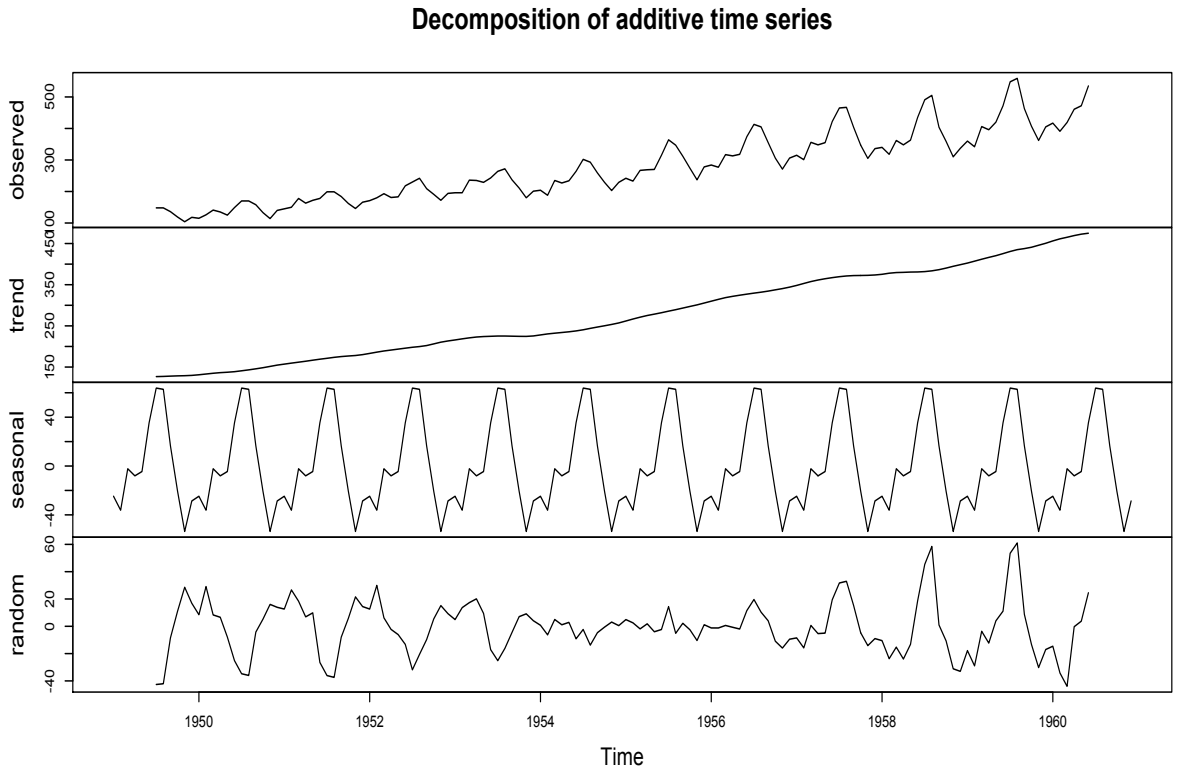
2. Le nombre de passagers : augmentation de moyenne + dispersion croissante

```
> x11()
> par(mfrow=c(2,1))
> plot(AirPassengers,type="l")
> plot(AirPassengers[1:50],type="l")
```






```
> plot(decompose(AirPassengers))
```

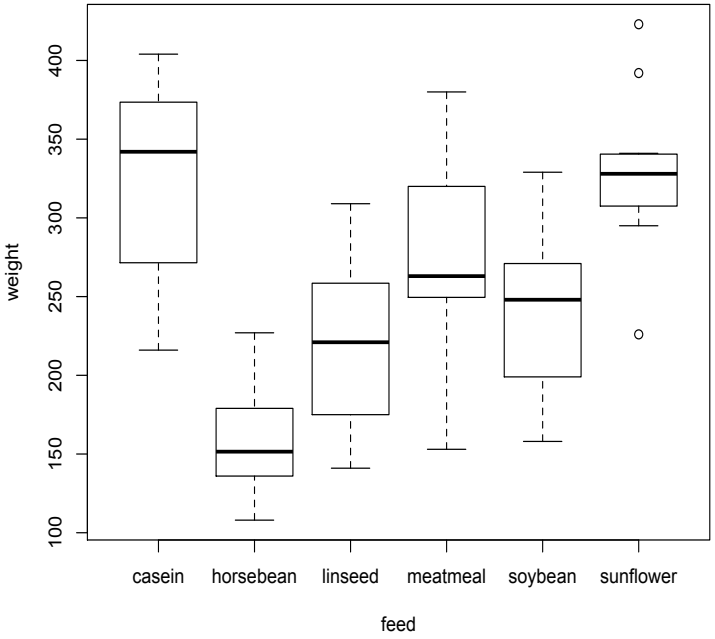


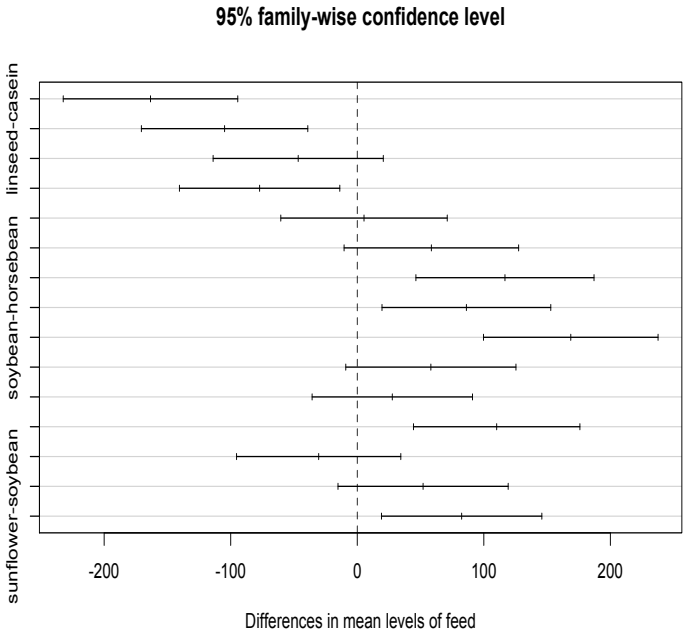
3. Masses de poulets en fonction de leur alimentation : analyse de la variance à un facteur

On veut tester s'il y a un effet du type de nourriture. Pour savoir quels sont les types de nourriture qui diffèrent des autres, on utilise la fonction TukeyHSD (Tukey's Honest Significant Difference). Les résultats sont montrés dans les figures et le tableau ci dessous. Tous les intervalles de confiance qui ne recoupent pas 0 révèlent des différences significatives d'effets.

Nourriture	Traduction
casein	caséine
horsebean	fève
linseed	graine de lin
meatmeal	farine animale
soybean	soja
sunflower	tournesol

```
> plot(weight~feed,data=chickwts)
> aov.chickwts=aov(weight~feed,data=chickwts)
> hsd.chickwts=TukeyHSD(aov.chickwts)
> plot(hsd.chickwts)
> hsd.chickwts
```






```
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = weight ~ feed, data = chickwts)
$feed
                           diff         lwr       upr     p adj
horsebean-casein    -163.383333 -232.346876 -94.41979 0.0000000
linseed-casein      -104.833333 -170.587491 -39.07918 0.0002100
meatmeal-casein      -46.674242 -113.906207  20.55772 0.3324584
soybean-casein       -77.154762 -140.517054 -13.79247 0.0083653
sunflower-casein       5.333333  -60.420825  71.08749 0.9998902
linseed-horsebean     58.550000  -10.413543 127.51354 0.1413329
meatmeal-horsebean   116.709091   46.335105 187.08308 0.0001062
soybean-horsebean     86.228571   19.541684 152.91546 0.0042167
sunflower-horsebean  168.716667   99.753124 237.68021 0.0000000
meatmeal-linseed      58.159091   -9.072873 125.39106 0.1276965
soybean-linseed       27.678571  -35.683721  91.04086 0.7932853
sunflower-linseed    110.166667   44.412509 175.92082 0.0000884
soybean-meatmeal     -30.480519  -95.375109  34.41407 0.7391356
sunflower-meatmeal    52.007576  -15.224388 119.23954 0.2206962
sunflower-soybean     82.488095   19.125803 145.85039 0.0038845
```

Chapitre 2

Régression linéaire simple

2.1 Linéarisation des principaux modèles utilisés en économétrie

Log-linéaire : $y = bx^a$ —> $\ln y = a \ln x + \ln b$

Exponentiel : $y = \exp(ax + b)$ —> $\ln y = ax + b$

Logarithmique : $y = a \ln x + b$

Hyperbolique : $y = \frac{a}{x - x_0} + y_0$

Parabolique : $y = ax^2 + bx + c$ —> $y = a(x - x_0)^2 + y_0$

Logistique : $y = y_{\min} + \frac{y_{\max} - y_{\min}}{1 + \exp(ax+b)}$ —> $\ln \frac{y_{\max} - y}{y - y_{\min}} = ax + b$

2.2 Principales lois et propriété de vecteurs gaussiens utilisées en modèle linéaire

Dans le cadre du modèle linéaire, nous allons essentiellement utiliser les lois de probabilités suivantes.

Loi du χ^2 à n degrés de liberté
Soient $X_1, \ldots, X_n$ n v. a. indép. de loi $\mathcal{N}(0, 1)$, alors

$$S = X_1^2 + \cdots + X_n^2$$

suit une loi du χ^2 à n degrés de liberté, notée $\chi^2(n)$.

Loi de Student à n degrés de liberté

La loi de Student à n degrés de liberté, notée $T(n)$, est la loi du quotient

$$T = \frac{N}{\sqrt{S/n}}$$

où N suit une loi $\mathcal{N}(0,1)$ et S suit une loi $\chi^2(n)$, N et S étant deux v. a. indép..

Loi de Fisher à n_1 et n_2 degrés de liberté

Soient S_1 et S_2 deux v. a. indép. de loi respectives $\chi^2(n_1)$ et $\chi^2(n_2)$. Alors le quotient

$$F = \frac{S_1/n_1}{S_2/n_2}$$

suit une loi de Fisher à n_1 et n_2 degrés de liberté, notée $F(n_1, n_2)$.

Voici une propriété de vecteurs gaussiens que nous utilisons le plus.

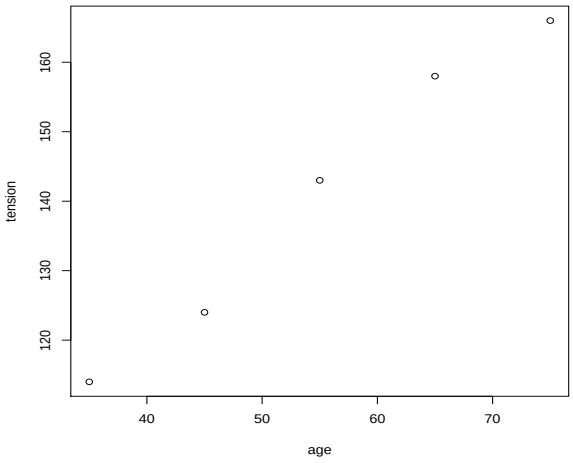
Théorème de Cochran

Soient E_1 et E_2 deux sous-espaces vectoriels orthogonaux de $E = \mathbb{R}^d$ de dimensions respectives k_1 et k_2 et soit Y un vecteur aléatoire de $\mathbb{R}^d$ de loi normale centrée isotrope de variance σ^2. Alors $P_{E_1}(Y)$ et $P_{E_2}(Y)$ sont deux v. a. gaussienne centrées indépendantes et $\|P_{E_1}(Y)\|^2$ (resp. $\|P_{E_2}(Y)\|^2$) est une loi $\sigma^2\chi^2(k_1)$ (resp. $\sigma^2\chi^2(k_2)$).

2.3 Un jeu de données

On considère un jeu de données (Y, X) où Y représente la tension artérielle et X représente l'âge.

Y	114	124	143	158	166
X	35	45	55	65	75



Commentaires du graphique :

– La tension artérielle augmente avec l'âge.

– Les points du graphe sont presque alignés.

2.4 Modèle et estimation

Nous proposons le modèle suivant :

$$Y_i = \mu + \beta X_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

Quatre présupposés du modèle : pour tous $i = 1, \ldots, n$

1. $\mathbb{E}(\varepsilon_i) = 0$
2. $\mathrm{Var}(\varepsilon_i) = \sigma^2$
3. ε_i sont i.i.d. de loi gaussienne.
4. X_i sont déterministes.

La *méthode des moindres carrés ordinaires* (MMCO) consiste à déterminer les valeurs μ et β en minimisant la *somme des carrés résiduelle* (SCR).

$$\mathrm{SCR}(\mu, \beta) := \sum_{i=1}^{n} (Y_i - (\mu + \beta X_i))^2.$$

Les estimateurs de MMCO sont les suivants

$$\hat{\beta} = \frac{\mathrm{Cov}(Y, X)}{\mathrm{Var}(X)}, \quad \hat{\mu} = \bar{Y} - \hat{\beta}\bar{X}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-2}$$

où

$$\mathrm{Cov}\,(Y,X)=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})(X_i-\bar{X}),\quad \mathrm{Var}\,(X)=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2,$$

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad \bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i,$$

$$\hat{Y}_i=\hat{\mu}+\hat{\beta}X_i\quad\text{et}\quad\hat{\varepsilon}_i=Y_i-\hat{Y}_i.$$

Remarque 1. Le coefficient $n-2$ de $\hat{\sigma}^2$ peut s'expliquer par la règle : le nombre de données (ici n) moins le nombre de paramètres du modèle (ici 2). Cette renormalisation a pour but d'obtenir un estimateur sans biais de σ^2, c.a.d. $\mathbb{E}\,(\hat{\sigma}^2)=\sigma^2$. En effet, $\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$ est une variable aléatoire de loi $\sigma^2\chi^2(n-2)$.

Notons Y^0 la nouvelle observation indépendante des premières observations vérifiant

$$Y^0=\mu+\beta X^0+\varepsilon^0\quad\text{et les quatre présupposés.}$$

Propriétés des estimateurs :

1. $\hat{\mu}\sim\mathcal{N}\left(\mu,\dfrac{\sigma^2\sum X_i^2}{n^2\mathrm{Var}\,(X)}\right)$
2. $\hat{\beta}\sim\mathcal{N}\left(\beta,\dfrac{\sigma^2}{n\mathrm{Var}\,(X)}\right)$
3. $\mathrm{Cov}\,(\hat{\mu},\hat{\beta})=-\dfrac{\sigma^2\bar{X}}{n\mathrm{Var}\,(X)}$
4. $\dfrac{\hat{\sigma}^2(n-2)}{\sigma^2}\sim\chi^2(n-2)$
5. $(\hat{\mu},\hat{\beta})$ et $\hat{\sigma}^2$ sont indépendantes.
6. $\mathrm{Var}\,(\hat{Y}^0)=\dfrac{\sigma^2}{n}\left(1+\dfrac{(X^0-\bar{X})^2}{\mathrm{Var}\,X}\right)$
7. Sous l'hypothèse nulle : $\beta=0$, la statistique $\hat{t}=\dfrac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$ suit la loi $T(n-2)$, où $\hat{\sigma}^2_{\hat{\beta}}=\dfrac{\hat{\sigma}^2}{n\mathrm{Var}\,X}$.
8. Sous l'hypothèse nulle : $\beta=0$, la statistique $\hat{F}=(n-2)\dfrac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}$ suit la loi $F(1,n-2)$.
9. L'intervalle de confiance au niveau $1-\alpha$ pour $\mathbb{E}\,(Y^0)$ est

$$\left[\hat{Y}^0-t_{n-2}(1-\alpha/2)\sqrt{\widehat{\mathrm{Var}}\,(\hat{Y}^0)},\quad\hat{Y}^0+t_{n-2}(1-\alpha/2)\sqrt{\widehat{\mathrm{Var}}\,(\hat{Y}^0)}\right].$$

10. L'intervalle de confiance au niveau $1-\alpha$ pour Y^0 est

$$\left[\hat{Y}^0 - t_{n-2}(1-\alpha/2)\sqrt{\widehat{\mathrm{Var}}\,(\hat{Y}^0) + \hat{\sigma}^2},\quad \hat{Y}^0 + t_{n-2}(1-\alpha/2)\sqrt{\widehat{\mathrm{Var}}\,(\hat{Y}^0) + \hat{\sigma}^2}\right]$$

où $\widehat{\mathrm{Var}}\,(\hat{Y}^0) = \dfrac{\hat{\sigma}^2}{n}\left(1 + \dfrac{(X^0 - \bar{X})^2}{\mathrm{Var}\,X}\right)$ et $t_{n-2}(1-\alpha/2)$ est le $1-\alpha/2$-quantile d'une loi de Student à $n-2$ degrés de liberté.

Sortie d'exemple traité par R

```
Call:
lm(formula = tension ~ age, data = Tens)

Residuals:
   1    2    3    4    5
 0.6 -3.2  2.0  3.2 -2.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
 (Intercept)  65.1000     5.8284   11.17 0.001538 **
 age           1.3800     0.1026   13.45 0.000889 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.246 on 3 degrees of freedom
Multiple R-squared: 0.9837,Adjusted R-squared: 0.9782
F-statistic: 180.8 on 1 and 3 DF,  p-value: 0.0008894
```

Les sorties de l'estimation des paramètres sont expliquées dans la table suivante ($k=1$).

	`Estimate`	`Std. Error`	`t value`	`Pr(>|t|)`	
`(Intercept)`	$\hat{\mu}$	$\hat{\sigma}_{\hat{\mu}}$	$\hat{t}_\mu = \hat{\mu}/\hat{\sigma}_{\hat{\mu}}$	$\mathbb{P}\left(\lvert t_{n-2}\rvert > \lvert\hat{t}_\mu\rvert\right)$	très significative
`age`	$\hat{\beta}$	$\hat{\sigma}_{\hat{\beta}}$	$\hat{t}_\beta = \hat{\beta}/\hat{\sigma}_{\hat{\beta}}$	$\mathbb{P}\left(\lvert t_{n-2}\rvert > \lvert\hat{t}_\beta\rvert\right)$	hautement significative

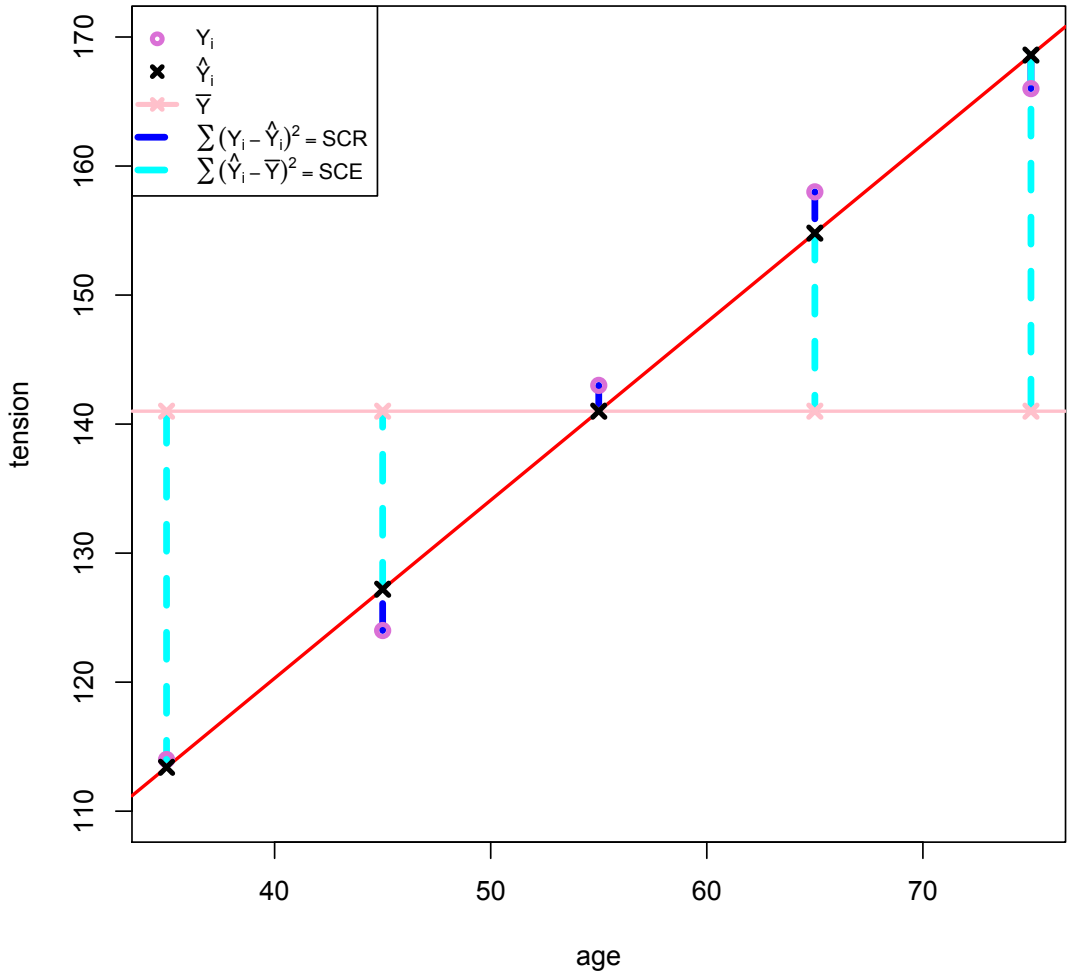
`Residual standard error`		$\hat{\sigma}$	
`Multiple R-squared`	$R^2 = \dfrac{\mathrm{SCE}}{\mathrm{SCT}}$	`Adjusted R-squared`	$R^2_{\text{ajusté}} = 1 - \dfrac{\mathrm{SCR}/(n-k-1)}{\mathrm{SCT}/(n-1)}$
`F-statistic`	$\hat{F} = (n-k-1)\dfrac{\mathrm{SCE}}{\mathrm{SCR}}$	`p-value`	$\mathbb{P}\left(F(k, n-k-1) > \hat{F}\right)$

Exercice 1. Réécrivez le modèle en utilisant les résultats de régression.

2.5 Analyse de la variance

Table d'analyse de la variance

Source	Somme des carrés	Degré de liberté	Carré moyen
expliquée	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 := \text{SCE}$	1	SCE
résiduelle	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 := \text{SCR}$	$n-2$	$\text{SCR}/(n-2)$
totale	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 := \text{SCT}$	$n-1$	$\text{SCT}/(n-1)$



Équation d'analyse de variance : $\text{SCT} = \text{SCE} + \text{SCR}$

Coefficient de détermination : $R^2 = \dfrac{\text{SCE}}{\text{SCT}}$

Coefficient de détermination ajusté : $R^2_{\text{ajusté}} = 1 - \dfrac{\text{SCR}/(n-k-1)}{\text{SCT}/(n-1)}$

Remarque 2. Ce coefficient est toujours compris entre 0 et 1. Il est un instrument de mesure de la qualité de l'ajustement, par le modèle linéaire et des données observées. Plus il est proche de 1, mieux cela vaut.

Sortie d'exemple traité par R

```
Analysis of Variance Table

Response: tension
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 1904.4 1904.40   180.8 0.0008894 ***
Residuals  3   31.6   10.53
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Les sorties de l'analyse de la variance sont expliquées dans la table suivante.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	SCE	SCE /1	$\hat{F}$	$\mathbb{P}\left(F(1, n-k-1) > \hat{F}\right)$
Residuals	$n-k-1$	SCR	SCR /$(n-k-1)$		

Exercice 2. Vérifiez l'équation d'analyse de variance. Le modèle permet il de bien expliquer les variations de la variable expliquée ? Que signifie-t-il le coefficient de détermination ?

Remarque 3. Pour sélectionner une spécification qui colle le mieux aux données, la comparaison directe entre deux R^2 du modèle différent peut être fortement trompeuse.

2.6 Calcul du pseudo-R^2

Supposons que vous hésitez entre une spécification linéaire et une spécification log-linéaire. La tentation est forte de comparer R^2 du modèle linéaire à R^2 du modèle log-linéaire linéarisé. En utilisant le jeu de données disposé, nous montrons que le procédé n'est qu'approximatif.

On modifie une observation de tension dans l'exemple "âge-tension", et refait la régression. Le coefficient R^2 obtenu est 0.9885.

Y	114	124	143	**156**	166
X	35	45	55	65	75

On modélise ensuite les données par le modèle log-linéaire, c'est-à-dire on applique la MMCO sur les données suivantes et estime les paramètres μ et β du modèle $\ln Y_i = \mu + \beta \ln X_i + \varepsilon_i$. Le coefficient R^2 obtenu est 0.9865.

$\ln Y$	4.74	4.82	4.96	5.05	5.11
$\ln X$	3.56	3.81	4.01	4.17	4.32

On ne peut pas conclure que le modèle linéaire simple est meilleur que le modèle log-linéaire, parce qu'on compare deux choses différentes : R^2 du modèle linéaire est la part expliquée de la variance de Y, tandis que celui du modèle log-linéaire est la part expliquée de la variance du logarithme de Y. Pour rendre la comparaison correcte, on calcule la part de la variance de Y qui est expliquée par le modèle log-linéaire de façon suivante.

$Y'_i = \exp(\widehat{\ln Y_i})$	On passe de la valeur prévue du logarithme de Y à la valeur correspondante pour Y.
$\mathrm{SCR}' = \sum_{i=1}^{5}(Y_i - Y'_i)^2$	La SCR est calculée à partir des écarts entre la valeur observée de Y et la valeur ainsi prévue.
Pseudo-$R^2 = 1 - \dfrac{\mathrm{SCR}'}{\mathrm{SCT}} = 0.9886$	La SCT doit être lu dans les sorties du modèle linéaire !

C'est ce coefficient qu'il faudrait en toute rigueur comparer au R^2 du modèle linéaire.

Remarque 4. Rien n'implique la moyenne des $\varepsilon'_i = Y_i - Y'_i$ soit nulle, ni dit que la variance de Y se décompose additivement en $\mathrm{Var}\,(Y') + \mathrm{Var}\,(\varepsilon')$. Il n'est donc plus possible de considérer que la mesure du pouvoir explicatif du modèle est complément à 1 de la mesure de son degré d'imprécision. C'est-à-dire l'équation d'analyse de variance n'est plus établie. Il faut alors faire un choix. Comme notre but est de voir si le modèle log-linéaire donne des prévisions plus précises des valeurs constatées de Y que le modèle linéaire, c'est l'aspect imprécision qu'il faut privilégier. Ceci conduit à la définition d'un pseudo R^2 pris égal au complément à 1 du degré d'imprécision.

Remarque 5. Sachez que le procédé de comparaison n'est qu'approximatif, une faible différence de R^2 n'a pas de signification.

2.7 Test sur la nullité de la pente β

On va tester l'hypothèse suivante.

$$H_0 : \beta = 0 \quad \text{contre} \quad H_1 : \beta \neq 0.$$

Deux tests vont permettre de décider quelle hypothèse choisir.

2.7.1 Test de Student

Sachant que $\hat{\beta}$ est un estimateur gaussien, et compte tenu du fait que $\hat{t} = \frac{\hat{\beta}-\beta}{\sqrt{\text{Var}\,(\hat{\beta})}}$ suit une loi de Student, la règle de décision à appliquer vient immédiatement

$$H_0 \Leftrightarrow |\hat{t}| = \frac{|\hat{\beta}|}{\sqrt{\text{Var}\,(\hat{\beta})}} \leq t_\alpha \quad \text{avec} \quad \alpha = \mathbb{P}\left(|t_{n-2}| > t_\alpha\right).$$

Dans les sorties de l'estimation des paramètres en R la valeur de "$Pr(>|t|)$" donne directement le résultat de ce test. C'est un indicateur de plausibilité de l'hypothèse H_0 qui prend une valeur comprise entre 0 et 1. Plus cette valeur est élevée, plus il sera raisonnable de retenir l'hypothèse H_0.

– significative dès lors que "$Pr(>|t|)" < 0.05$
– très significative dès lors que "$Pr(>|t|)" < 0.01$
– hautement significative dès lors que "$Pr(>|t|)"< 0.001$

2.7.2 Test de Fisher

Ce test se fonde sur l'équation d'analyse de variance

$$\text{SCT} = \text{SCE} + \text{SCR}\,.$$

Sous les hypothèses de la MMCO, on peut démontrer que

$$\mathbb{E}\left(\frac{\text{SCR}}{n-2}\right) = \sigma^2.$$

On peut démontrer aussi que

$$H_0 : \beta = 0 \Leftrightarrow \mathbb{E}\,(\text{SCE}\,) = \sigma^2 \quad \text{et que} \quad H_1 : \beta \neq 0 \Leftrightarrow \mathbb{E}\,(\text{SCE}\,) > \sigma^2.$$

In en résulte l'équivalence absolue des deux couples d'hypothèses suivantes.

$$H_0 : \beta = 0 \Leftrightarrow H_0 : \frac{\mathbb{E}\,(\text{SCE}\,)(n-2)}{\mathbb{E}\,(\text{SCR}\,)} = 1$$

$$H_1 : \beta \neq 0 \Leftrightarrow H_1 : \frac{\mathbb{E}\,(\text{SCE}\,)(n-2)}{\mathbb{E}\,(\text{SCR}\,)} > 1$$

Prenons la statistique de test

$$\hat{F} = (n-2)\frac{\text{SCE}}{\text{SCR}}.$$

On peut démontrer que cette statistique suit une loi de Fisher à $(1, n-2)$ degrés de liberté. La règle de décision s'écrit finalement

$$H_0 \Leftrightarrow \hat{F} \leq F_\alpha \quad \text{avec} \quad \alpha = \mathbb{P}\left(F(1, n-2) > F_\alpha\right).$$

Dans les sorties de l'analyse de la variance en R la valeur de $Pr(> F)$ donne directement le résultat de ce test. C'est un indicateur de plausibilité de l'hypothèse H_0 qui prend une valeur comprise entre 0 et 1. Plus cette valeur est élevée, plus il sera raisonnable de retenir l'hypothèse H_0.

Remarque 6. Équivalence des tests de Student et de Fisher :

$$\hat{t}^2 = \hat{F}$$

Il n'y a aucune déférence entre la règle de décision de Student et celle de Fisher dans le cas du modèle à une variable explicative.

2.8 Rappel aux tests statistiques

Un test statistique permet de décider entre deux hypothèses H_0 et H_1. Cette décision se fera à partir d'une réalisation d'un échantillon. On associe à un test un niveau α, ou risque de première espèce (généralement entre 1% et 10%). Une fois que le niveau α est fixé on peut en déduire la région de rejet.

2.8.1 Construction d'un test

1. Choix du niveau α

Un choix standard est $\alpha = 5\%$. Pour savoir plus sur cette norme officieuse, voir la section 2.4 "La canonisation du 5%" dans [2].

2. Choix des hypothèses H_0 et H_1

On veut tester l'appartenance du paramètre θ à un ensemble de valeurs Θ_0. On forme donc le test

$$H_0 : \theta \in \Theta_0 \quad \text{contre} \quad H_1 : \theta \in \Theta_1,$$

où $\Theta_0 \cap \Theta_1 = \emptyset$. Une hypothèse est dite simple si elle est associée à un singleton, c'est-à-dire $H_0 : \theta = \theta_0$ ou $H_1 : \theta = \theta_1$, sinon elle sera dite multiple, ou bien composite. Si H_1 est multiple de la forme $\theta > \theta_0$ ou $\theta < \theta_0$, on parlera de test unilatéral. Si H_1 est multiple de la forme $\theta \neq \theta_0$ on parlera de test bilatéral.

3. Choix de la statistique de test

La statistique est une quantité calculée à partir d'un échantillon qui suit une loi connue sous l'hypothèse H_0. On note T "la statistique théorique" – la variable aléatoire qui suit la loi de statistique choisie, $\hat{T}$ "la statistique empirique" – la statistique calculée à partir de la réalisation de l'échantillon.

4. Détermination de la région de rejet

La région de rejet, notée W, est l'ensemble de valeur auquel la statistique choisie appartient sous l'hypothèse H_0 avec la probabilité égale à α, i.e.

$$\mathbb{P}\left(T \in W \mid H_0\right) = \alpha.$$

Autrement dit si H_0 est vraie la statistique T a peu de chance d'être dans W. Donc si la statistique calculée $\hat{T}$ fait partie de W, on a la raison suffisante pour rejeter H_0, d'où vient le nom "la région de rejet".

5. Conclusion pour la réalisation de l'échantillon

On calcule la statistique $\hat{T}$ et compare avec la région de rejet ; La décision est prise de façon suivante.

$$\begin{cases} \text{rejet de } H_0 & \text{si } \hat{T} \in W, \\ \text{acceptation de } H_0 & \text{si } \hat{T} \notin W. \end{cases}$$

Remarque 7. Un test a le plus souvent pour but de rejeter H_0 plutôt que de choisir entre deux hypothèses, puisque **accepter H_0 ne veut pas dire que cette hypothèse est vraie mais qu'il n'y pas de raison suffisante pour la rejeter. Rejeter H_0 est donc beaucoup riche en information que l'accepter**. C'est pour cela que dans le cas de rejet on dit que le test est significatif. Les deux hypothèses ne sont donc pas interchangeables.

Un exemple simple est présenté ci dessous pour illustrer la construction d'un test et en séduire l'intérêt de la notion P-value.

2.8.2 Exemple simple

On dispose d'un échantillon gaussien $X_1, \ldots, X_n$ de loi $\mathcal{N}(\mu, \sigma^2)$ avec μ inconnu et on veut réaliser le test suivant,

$$H_0 : \mu = m \quad \text{contre} \quad H_1 : \mu \neq m,$$

pour une valeur m donnée. On choisit intuitivement [1] la statistique de test $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. On sait que sous H_0, $\bar{X}$ est distribuée suivant une loi gaussienne $\mathcal{N}(\mu, \sigma^2/n)$, ainsi $\frac{\sqrt{n}}{\sigma}(\bar{X} -$

1. On peut bien sûr choisir une autre statistique, par exemple X_1. Qu'obtient-on ?

$m) \sim \mathcal{N}(0,1)$. Soit un niveau $\alpha = 5\%$. On trouve dans la table de loi gaussienne centrée et réduite le quantile $q_{97.5\%} = 1.96$, i.e.

$$\mathbb{P}\,(\frac{\sqrt{n}}{\sigma}(\bar{X} - m) < 1.96 \,|\, H_0) = 0.975.$$

Puisque le test est bilatéral, on déduit

$$\mathbb{P}\,(\frac{\sqrt{n}}{\sigma}|\bar{X} - m| \geq 1.96 \,|\, H_0) = 0.05,$$

ainsi,

$$\mathbb{P}\,(\bar{X} \in (-\infty, m - 1.96\sigma/\sqrt{n}) \cup (m + 1.96\sigma/\sqrt{n}, \infty) \,|\, H_0) = 0.05.$$

La région de rejet est donc

$$W = (-\infty, m - 1.96\sigma/\sqrt{n}) \cup (m + 1.96\sigma/\sqrt{n}, \infty).$$

Notons $\hat{\bar{X}}$ la moyenne empirique calculée à partir d'une réalisation de l'échantillon. La décision est la suivante.

$$\begin{cases} \text{rejet de } \ H_0 & \text{si } \ \hat{\bar{X}} < m - 1.96\sigma/\sqrt{n} \ \text{ ou } \ \hat{\bar{X}} > m + 1.96\sigma/\sqrt{n}, \\ \text{acceptation de } \ H_0 & \text{si } \ m - 1.96\sigma/\sqrt{n} \leq \hat{\bar{X}} \leq m + 1.96\sigma/\sqrt{n}. \end{cases}$$

Cela veut dire qu'on rejettera l'hypothèse $H_0 : \mu = m$ si la moyenne observée est "loin" de m. Le niveau α est la probabilité de commettre une erreur qu'on rejette H_0 alors qu'elle est vraie. Autrement dit le risque de rejeter à tort l'hypothèse nulle est 5%.

Remarquons que la région de rejet est liée au choix du niveau α. Pour un niveau $\alpha = 10\%$ on obtient la région de rejet suivante

$$W = (-\infty, m - 1.64\sigma/\sqrt{n}) \cup (m + 1.64\sigma/\sqrt{n}, \infty),$$

qui inclut celle de $\alpha = 5\%$. Ainsi pour tous

$$\hat{\bar{X}} \in (m - 1.96\sigma/\sqrt{n}, m - 1.64\sigma/\sqrt{n}) \cup (m + 1.64\sigma/\sqrt{n}, m + 1.96\sigma/\sqrt{n}),$$

l'hypothèse H_0 sera rejetée si $\alpha = 10\%$, mais acceptée si $\alpha = 5\%$. C'est-à-dire, la décision du test dépend non seulement des observations, mais aussi du niveau α. Plus le niveau choisi est petit, plus il est difficile d'obtenir le résultat significatif.

Rappelons que le niveau α est en fait le risque seuil, en dessous duquel on est prêt à rejeter H_0. Un risque de 5% veut dire que dans 5% des cas quand H_0 est vraie, l'expéri-

mentateur se trompera et la rejettera. Mais le choix du seuil à employer dépendra de la certitude désirée et de la vraisemblance des alternatives. On souhaite donc d'obtenir le résultat qui est indépendant du risque seuil et permet de choisir le risque seuil a posteriori.

2.8.3 La P-value

La P-value ou niveau de signification est la valeur critique de α qui fait basculer le résultat du test. Elle dépend uniquement de la réalisation de l'échantillon et permet de faire la décision avec le niveau α choisi arbitrairement. On a

$$\begin{cases} \text{rejet de } \ H_0 & \text{si } \ \alpha > \text{P-value}, \\ \text{acceptation de } \ H_0 & \text{si } \ \alpha < \text{P-value}. \end{cases}$$

La P-value est en fait la valeur la plus petite de α qui permet de rejeter H_0. Si la P-value est plus grande que α choisi a priori, le test est non concluant, ce qui revient à dire que l'on ne peut rien affirmer. Prenons l'exemple du test bilatéral présenté précédemment, la P-value peur être calculée,

$$\text{P-value} = \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}|\bar{X} - m| \geq \frac{\sqrt{n}}{\sigma}|\hat{\bar{X}} - m| \,\middle|\, H_0 \right).$$

C'est la probabilité, en supposant que H_0 est vraie, d'obtenir une valeur de la variable de décision $|\bar{X} - m|$ au moins aussi grande que la valeur de la statistique que l'on a obtenue avec notre échantillon $|\hat{\bar{X}} - m|$. Puisque sous H_0, on a $\frac{\sqrt{n}}{\sigma}(\bar{X} - m) \sim \mathcal{N}(0,1)$. En notant $\mathcal{N}(0,1)$ une variable aléatoire gaussienne standard, on a

$$\text{P-value} = \mathbb{P}\left(|\mathcal{N}(0,1)| \geq \frac{\sqrt{n}}{\sigma}|\hat{\bar{X}} - m| \right).$$

Chapitre 3

Régression linéaire multiple

Notre but est d'étudier la relation de détermination de Y par k variables explicatives $X^{(1)}, \ldots, X^{(k)}$,

$$Y_i = \mu + \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \cdots + \beta_k X_i^{(k)} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Il s'en suit l'écriture matricielle suivante

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1^{(1)} & X_1^{(2)} & \cdots & X_1^{(k)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n^{(1)} & X_n^{(2)} & \cdots & X_n^{(k)} \end{pmatrix} \begin{pmatrix} \mu \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

ou encore

$$Y = X\theta + \varepsilon \quad \text{avec} \quad \theta = \begin{pmatrix} \mu \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \text{et} \quad X = \begin{pmatrix} 1 & X_1^{(1)} & X_1^{(2)} & \cdots & X_1^{(k)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n^{(1)} & X_n^{(2)} & \cdots & X_n^{(k)} \end{pmatrix}.$$

Les quatre présupposés s'expriment

$$\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$$

où I_n est la matrice identité d'ordre n. On peut encore les formuler

$$Y \sim \mathcal{N}(X\theta, \sigma^2 I_n).$$

3.1 Le modèle linéaire en 4 formules

Les 4 formules fondamentales sont issues de la minimisation en θ de la somme des carrés résiduelle (SCR), somme qui peut s'écrire matriciellement sous la forme

$$\mathrm{SCR}\,(\theta) = \|Y - X\theta\|^2 = (Y - X\theta)'(Y - X\theta).$$

On a alors

1. $\hat{\theta} = (X'X)^{-1}X'Y$
2. $\mathbb{E}\,(\hat{\theta}) = \theta$
3. $\mathrm{Var}\,(\hat{\theta}) = \sigma^2(X'X)^{-1}$
4. $\mathrm{SCR}\,(\hat{\theta}) = \|Y - \hat{Y}\|^2$ est une variable aléatoire indépendante de $\hat{\theta}$ et suit une loi $\sigma^2\chi^2(n-k-1)$.

Nous pouvons utiliser le théorème de Pythagore et le théorème de Cochran pour démontrer la 4ème formule.

Dans la suite, notons $\hat{Y} = X\hat{\theta}$ et $\hat{\varepsilon} = Y - \hat{Y}$. Nous avons les propriétés suivantes :

$$\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2(X'X)^{-1}) \quad \text{et} \quad \hat{Y} \sim \mathcal{N}(X\theta, \sigma^2 X(X'X)^{-1}X').$$

Étant donnée X^0 matrice $m \times (k+1)$ de valeurs des explicatives, pour m observations indépendantes des premières observations et vérifiant

$$Y^0 = X^0\theta + \varepsilon^0 \quad \text{et} \quad \varepsilon^0 \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_m).$$

En notant $\hat{Y}^0 = X^0\hat{\theta}$, nous avons

$$\hat{Y}^0 \sim \mathcal{N}(X^0\theta, \sigma^2 X^0(X'X)^{-1}X^{0\prime}) \quad \text{et} \quad \hat{Y}^0 - Y^0 \sim \mathcal{N}(\mathbf{0}, \sigma^2(X^0(X'X)^{-1}X^{0\prime} + I_m)).$$

Le tableau ci dessous récapitule les bandes de confiance et de prédiction pour le modèle multiple.

Objectif	Intervalle de confiance
$\mathbb{E}\,(Y^0)$	$\hat{Y}^0 \pm t_{n-k-1}(1-\alpha/2)\sqrt{diag(\hat{\sigma}^2 X^0(X'X)^{-1}X^{0\prime})}$
Y^0	$\hat{Y}^0 \pm t_{n-k-1}(1-\alpha/2)\sqrt{diag(\hat{\sigma}^2(X^0(X'X)^{-1}X^{0\prime} + I_m))}$

3.1.1 Transformations linéaires sur les variables

Notons

$$\tilde{Y}_i = aY_i + b, \quad \tilde{X}_i^{(j)} = a_j X_i^{(j)} + b_j, \quad j = 1, \ldots, n,$$

et supposons que la relation de détermination de $\tilde{Y}$ par $\tilde{X}^{(j)}$ s'écrit de façon suivante

$$\tilde{Y}_i = \tilde{\mu} + \tilde{\beta}_1 \tilde{X}_i^{(1)} + \tilde{\beta}_2 \tilde{X}_i^{(2)} + \cdots + \tilde{\beta}_k \tilde{X}_i^{(k)} + \tilde{\varepsilon}_i, \quad i = 1, \ldots, n.$$

Il est facile de montrer que la relation entre les paramètres du modèle initial et ceux du modèle transformé est la suivante

$$\mu = (\tilde{\mu} - b + \tilde{\beta}_1 b_1 + \cdots + \tilde{\beta}_k b_k)/a$$

et

$$\beta_j = \tilde{\beta}_j a_j / a.$$

Remarque 8. Le centrage ne modifie pas la valeur des coefficients estimés des variables. Par contre le terme constant estimé du modèle en variables centrées n'est pas le même que celui du modèle initial. Il est d'ailleurs forcément égal à 0.

Exercice 3. Démontrer la remarque précédente.

Notons

$$\tilde{Y}_i = Y_i - \bar{Y}, \quad \tilde{X}_i^{(j)} = X_i^{(j)} - \overline{X^{(j)}}, \quad j = 1, \ldots, k, \quad i = 1, \ldots, n.$$

D'après la remarque 8 nous avons

$$\tilde{Y}_i = \beta_1 \tilde{X}_i^{(1)} + \beta_2 \tilde{X}_i^{(2)} + \cdots + \beta_k \tilde{X}_i^{(k)} + \tilde{\varepsilon}_i, \quad i = 1, \ldots, n.$$

L'intérêt du travail en variable centrées est d'obtenir une matrice $\tilde{X}'\tilde{X}$ plus simple. On a

$$\tilde{X}'\tilde{X} = \begin{pmatrix} n & 0 \\ 0 & n\mathrm{Var}\,(X) \end{pmatrix} \tag{1}$$

où $\mathrm{Var}\,(X)$ est la matrice des variances covariances des variables $X^{(j)}$.

Exercice 4. Démontrer l'expression (1).

Le travail en variables centrées-réduites

Notons cette fois ci

$$\tilde{Y}_i = (Y_i - \bar{Y})/\sigma_Y, \quad \tilde{X}_i^{(j)} = (X_i^{(j)} - \overline{X^{(j)}})/\sigma_{X^{(j)}}, \quad j = 1, \ldots, k, \quad i = 1, \ldots, n.$$

Comme les variables sont centrées, l'estimation du terme constant est nulle. Nous ne nous intéressons qu'aux coefficients des variables et à la partie de la matrice $\tilde{X}'\tilde{X}$ qui les concerne. Appelons $\hat{\tilde{\theta}}$ le vecteur de ces coefficients estimés, i.e.,

$$\hat{\tilde{\theta}} = \begin{pmatrix} \hat{\tilde{\beta}}_1 \\ \vdots \\ \hat{\tilde{\beta}}_k \end{pmatrix}.$$

Il est facile de montrer que, dans ce cas, la formule d'estimation des coefficients peut s'écrire

$$\hat{\tilde{\theta}} = \mathrm{Cor}\,(X)^{-1}\mathrm{Cor}\,(Y, X)$$

où $\mathrm{Cor}\,(X)$ est la matrice des coefficients de corrélation linéaire des variables explicatives prises deux à deux, et $\mathrm{Cor}\,(Y, X)$ est le vecteur des coefficients de corrélation linéaire simple entre Y et chacune des variables $X^{(j)}$.

Remarque 9. La valeur des coefficients d'une régression dépend de l'unité de mesure des variables. L'intérêt de la réduction des variables est d'éliminer l'influence, sur la valeur des coefficients, des conventions d'unités qui portent sur les mesures des variables, et d'adopter pour toutes les variables une échelle commune libellée en unité d'écart-type de chaque variable.

3.2 La valeur explicative du modèle dans les données

Coefficient de détermination multiple : $R^2 = \dfrac{\mathrm{SCE}}{\mathrm{SCT}}$

Coefficient de détermination partielle : Le coefficient de détermination partielle associé à une variable quelconque est la part de la variance laissée inexpliquée par les autres variables qui est expliquée grâce à l'introduction de la variable étudiée.

Calcul

– Une première régression en l'absence de la variable étudiée $X^{(j)}$. On a

$$\mathrm{SCT}_1 = \mathrm{SCE}_1 + \mathrm{SCR}_1 \quad \text{(modèle 1)}.$$

– Une seconde régression en prenant toutes les variables, la variable étudiée $X^{(j)}$ comprise. On a

$$\mathrm{SCT}_2 = \mathrm{SCE}_2 + \mathrm{SCR}_2 \quad \text{(modèle 2)}.$$

Le coefficient de détermination partielle associé à $X^{(j)}$ est défini par la formule suivante

$$R^2_{(j)} = \frac{\mathrm{SCE}_2 - \mathrm{SCE}_1}{\mathrm{SCR}_1}.$$

SCR $_1$: ce qui n'est pas expliqué par la première régression (en l'absence de $X^{(j)}$).

SCE $_2$ **− SCE** $_1$: le supplément d'explication apporté par $X^{(j)}$.

Remarque 10. Ce coefficient est compris entre 0 et 1. Une valeur nulle indique que la variable $X^{(j)}$ n'apporte rien de plus à l'explication de Y. Une valeur unitaire indique, au contraire, que cette varaible explique tout ce qui restait à expliquer.

Remarque 11. D'ailleurs, on a aussi

$$R^2_{(j)} = \frac{\mathrm{SCR}_1 - \mathrm{SCR}_2}{\mathrm{SCR}_1} = \frac{R^2_{\text{modèle 2}} - R^2_{\text{modèle 1}}}{1 - R^2_{\text{modèle 1}}}.$$

Une valeur unitaire de $R^2_{(j)}$ implique que le coefficient de détermination du modèle 2, $R^2_{\text{modèle 2}}$, est égal à 1. Si ceci se produit les coefficients de détermination partielle de toutes les varaibles seront égaux à 1.

3.3 Les tests

3.3.1 Le test de significativité globale

H_0 : aucune des variables n'a d'action sur $Y \Leftrightarrow \forall j \in \{1, \ldots, k\}, \beta_j = 0$.

H_1 : au moins une des variables a une action sur $Y \Leftrightarrow \exists j \in \{1, \ldots, k\}$, tel que $\beta_j \neq 0$.

$$H_0 \Leftrightarrow \hat{F} = \frac{n-k-1}{k}\frac{\mathrm{SCE}}{\mathrm{SCR}} \leq F_\alpha \quad \text{avec} \quad \alpha = \mathbb{P}\left(F(k, n-k-1) > F_\alpha\right)$$

3.3.2 Le test de significativité partielle

H_0 : $X^{(j)}$ n'a pas d'action sur $Y \Leftrightarrow \beta_j = 0$.

H_1 : $X^{(j)}$ a une action sur $Y \Leftrightarrow \beta_j \neq 0$.

Test de Student

$$H_0 \Leftrightarrow \hat{t_j} = \frac{|\hat{\beta}_j|}{\sqrt{\mathrm{Var}\left(\hat{\beta}_j\right)}} \leq t_\alpha \quad \text{avec} \quad \alpha = \mathbb{P}\left(|t_{n-k-1}| > t_\alpha\right)$$

Test du Fisher partiel

$$H_0 \Leftrightarrow \hat{F_j} = \frac{\mathrm{SCE}_2 - \mathrm{SCE}_1}{\mathrm{SCR}_2}(n-k-1) \leq F_\alpha \quad \text{avec} \quad \alpha = \mathbb{P}\left(F(1, n-k-1) > F_\alpha\right)$$

La section 3.5 présente les paradoxes possibles de la significativité partielle et le problème de la colinéarité statistique.

3.4 Autres utilisations du test de Student

Il peut être intéressant de tester la position des paramètres par rapport à des valeurs particulières, ce qu'autorise le test de Student, comme l'indique le tableau ci dessous.

test unilatéral droit	$H_0 : \theta = \theta_0$ contre $H_1 : \theta > \theta_0$ $W = \left\{ \hat{t} = \frac{\hat{\theta}-\theta_0}{\hat{\sigma}_{\hat{\theta}}} > t_\alpha \right\}$ avec $\alpha = \mathbb{P}\left(t_{n-k-1} > t_\alpha\right)$
test unilatéral gauche	$H_0 : \theta = \theta_0$ contre $H_1 : \theta < \theta_0$ $W = \left\{ \hat{t} = \frac{\hat{\theta}-\theta_0}{\hat{\sigma}_{\hat{\theta}}} < t_\alpha \right\}$ avec $\alpha = \mathbb{P}\left(t_{n-k-1} < t_\alpha\right)$
test bilatéral	$H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ $W = \left\{ \hat{t} = \frac{\lvert\hat{\theta}-\theta_0\rvert}{\hat{\sigma}_{\hat{\theta}}} > t_\alpha \right\}$ avec $\alpha = \mathbb{P}\left(\lvert t_{n-k-1}\rvert > t_\alpha\right)$

La notation θ désigne le coefficient β_j de la variable $X^{(j)}$ ou le terme constant μ du modèle linéaire estimé par la MMCO.

Remarque 12. Par symétrie de la loi de Student, on a

$$\alpha = \mathbb{P}\left(t_{n-k-1} > t_\alpha\right) = \mathbb{P}\left(t_{n-k-1} < -t_\alpha\right).$$

Ainsi dans le test unilatéral gauche, la région de rejet peut s'écrire $W = \left\{ \hat{t} = \frac{\hat{\theta}-\theta_0}{\hat{\sigma}_{\hat{\theta}}} < -t_\alpha \right\}$ avec $\alpha = \mathbb{P}\left(t_{n-k-1} > t_\alpha\right)$.

3.5 Colinéarité statistique

Imaginons qu'un modèle soit globalement significatif. Est-il possible que les tests de significativité partielle se révèlent tous négatifs ? La situation serait paradoxale : globalement significatif veut dire qu'au moins une des variables a une action sur Y ; si tous les tests partiels sont négatifs, cela signifie qu'on ne trouve aucune variable dont l'action soit significative. Cette apparente contradiction peut toutefois se rencontrer.

Le problème vient de la colinéarité statistique qui peut exister malencontreusement entre les variables explicatives. Un exemple nous suffira à illustrer ce fait. Nous avons recueilli deux séries d'observations sur des variables Y, X_1 et X_2. Le nombre d'observations

est le même dans les deux cas ($n = 10$). Les données figurent dans les tableaux ci-dessous.

Première série d'observations

Y	22.73	21.36	24.09	26.89	24.95	32.88	33.54	31.39	36.85	30.69
X_1	10	11	12	13	14	15	16	17	18	19
X_2	14	12	11	14	11	18	18	15	15	12

Sorties d'estimation traitée par R

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.6757     3.2989  -0.811 0.444022
X1            1.2413     0.1776   6.990 0.000213 ***
X2            0.9439     0.2082   4.533 0.002689 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.539 on 7 degrees of freedom
Multiple R-squared: 0.9327,Adjusted R-squared: 0.9134
F-statistic: 48.49 on 2 and 7 DF,  p-value: 7.918e-05
```

Deuxième série d'observations

Y	21.99	21.37	24.72	27.16	30.60	31.52	33.35	38.21	33.55	40.29
X_1	10	11	12	13	14	15	16	17	18	19
X_2	12	11	12	13	16	15	16	17	18	21

Sorties d'estimation traitée par R

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1012     3.1266   0.032   0.9751
x1            1.3180     0.6795   1.940   0.0936 .
x2            0.7327     0.6545   1.119   0.2999
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.896 on 7 degrees of freedom
Multiple R-squared: 0.9327,Adjusted R-squared: 0.9134
```

```
F-statistic: 48.48 on 2 and 7 DF,  p-value: 7.921e-05
```

Les deux modèles sont donc tous globalement très significatifs. Les deux coefficients du modèle 1 sont très significativement différents de 0, tandis qu'aucune des deux variables du modèle 2 ne semble avoir d'action significative sur Y au seuil 5%. Ce paradoxe du deuxième modèle s'explique par une forte colinéarité entre X_1 et X_2. Cette observation n'est pas casuelle : il est facile de montrer en effet que la covariance entre variables explicatives accroît la variance de chacun des coefficients de ces variables.

Puisque nous nous intéressons seulement aux coefficients des variables, le plus simple est de raisonner en variables centrées réduites. On déduit de (1) le bloc de la matrice $\tilde{X}'\tilde{X}$ qui concerne les coefficients des variables est $n\mathrm{Cor}\,(X)$. Il s'ensuit que la matrice des variances covariances de $\hat{\tilde{\theta}}$ s'exprime

$$\mathrm{Var}\,(\hat{\tilde{\theta}}) = \mathrm{Cor}\,(X)^{-1}\sigma^2/n.$$

Plaçons nous dans le cas où il n'y a que deux variables explicatives et appelons ρ leur coefficient de corrélation linéaire. Alors on a

$$\mathrm{Cor}\,(X) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

ainsi

$$\mathrm{Cor}\,(X)^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

Il s'ensuit que les termes de la diagonale de $\mathrm{Var}\,(\hat{\tilde{\theta}})$ sont tous de la forme

$$\frac{\sigma^2}{n(1-\rho^2)}.$$

Plus grande sera la corrélation entre nos deux variables explicatives, plus grande sera la variance de leurs coefficients estimés. La colinéarité statistique entre variables explicatives ne fait donc pas que rendre covariantes les estimations des coefficients des variables, elle en augmente les variances, ce qui rend plus incertaines ces estimations et peut conduire à les faire apparaître non-significativement différentes de 0.

Remarque 13. Lorsqu'une variable présente dans une régression n'est pas significative, cela ne veut pas nécessairement dire qu'elle n'a pas d'action : son action peut être réelle mais n'être pas prouvée du fait d'une colinéarité statistique avec les autres variables également présentes dans cette régression.

3.6 Algorithmes de construction de modèle ne comprenant que des variables significatives

3.6.1 Régression descendante

Étape 1 : On part du modèle avec toutes les variables exogènes.

Étape 2 : On écarte la moins significative des variables non-significatives. C'est-à-dire on teste la significativité de toutes les variables. Si elle est établie, alors **ARRÊT**. Si non, on sort la variable dont le $\hat{t}$ est le plus faible.

Étape 3 : On recommence Étape 2 jusqu'à l'obtention d'un modèle qui ne contienne que des variables significatives.

3.6.2 Régression ascendante

Étape 1 : On part de rien. On sélectionne la variable qui contribue le mieux à expliquer la variable Y, c.a.d. celle dont le coefficient de corrélation est le plus élevé. On vérifie avec un test de Fisher si l'apport d'explication de cette variable est significatif. Si non, alors **ARRÊT**.

Étape 2 : On sélectionne parmi les variables restantes, celle qui contribuerait le mieux à expliquer le résidu estimé, c.a.d. celle dont le coefficient de détermination partielle est le plus élevé. On vérifie par un test du Fisher partiel que le supplément d'explication est significatif. Si ce n'est pas le cas, alors **ARRÊT**.

Étape 3 : On recommence Étape 2 jusqu'à la situation où aucune variable restante n'apporte un supplément significatif d'explication des variations de Y.

3.7 Étude de cas : AirPassengers

Notons Y_i les nombres de passagers mensuels, t_i la variable de temps. On commence par les trois modèles suivants.

1. $Y_i = \mu + \beta t_i + \varepsilon_i$
2. $\log Y_i = \mu + \beta t_i + \varepsilon_i$
3. $\log Y_i = \mu + \beta_1 t_i + \beta_2 (t_i - \bar{t})^2 + \varepsilon_i$

Les résultats de l'estimation et du test sont résumés dans le tableau suivant.

	`mod1`	`mod2`	`mod3`
Code en R	`AP~temps`	`log(AP)~temps`	`log(AP)~temps+(temps-mean(temps))^2`
Modèle	$Y_i = \mu + \beta t_i + \varepsilon_i$	$\log Y_i = \mu + \beta t_i + \varepsilon_i$	$\log Y_i = \mu + \beta_1 t_i + \beta_2 (t_i - \bar{t})^2 + \varepsilon_i$
Estimation	$\hat{\mu} = -62056, \hat{\beta} = 31.89$	$\hat{\mu} = -230, \hat{\beta} = 0.12$	$\hat{\mu} = -230, \hat{\beta}_1 = 0.12, \hat{\beta}_2 = -0.0032$
R^2	0.8536	0.9015(0.8535)	0.9074(0.8613)
$(\hat{Y}_i, \hat{\varepsilon}_i)$	non constante	parabolique	pas de remarque
QQ-plot	courbure évidente	presque aligné	presque aligné
Shapiro Test	10^{-5}	0.09	0.05

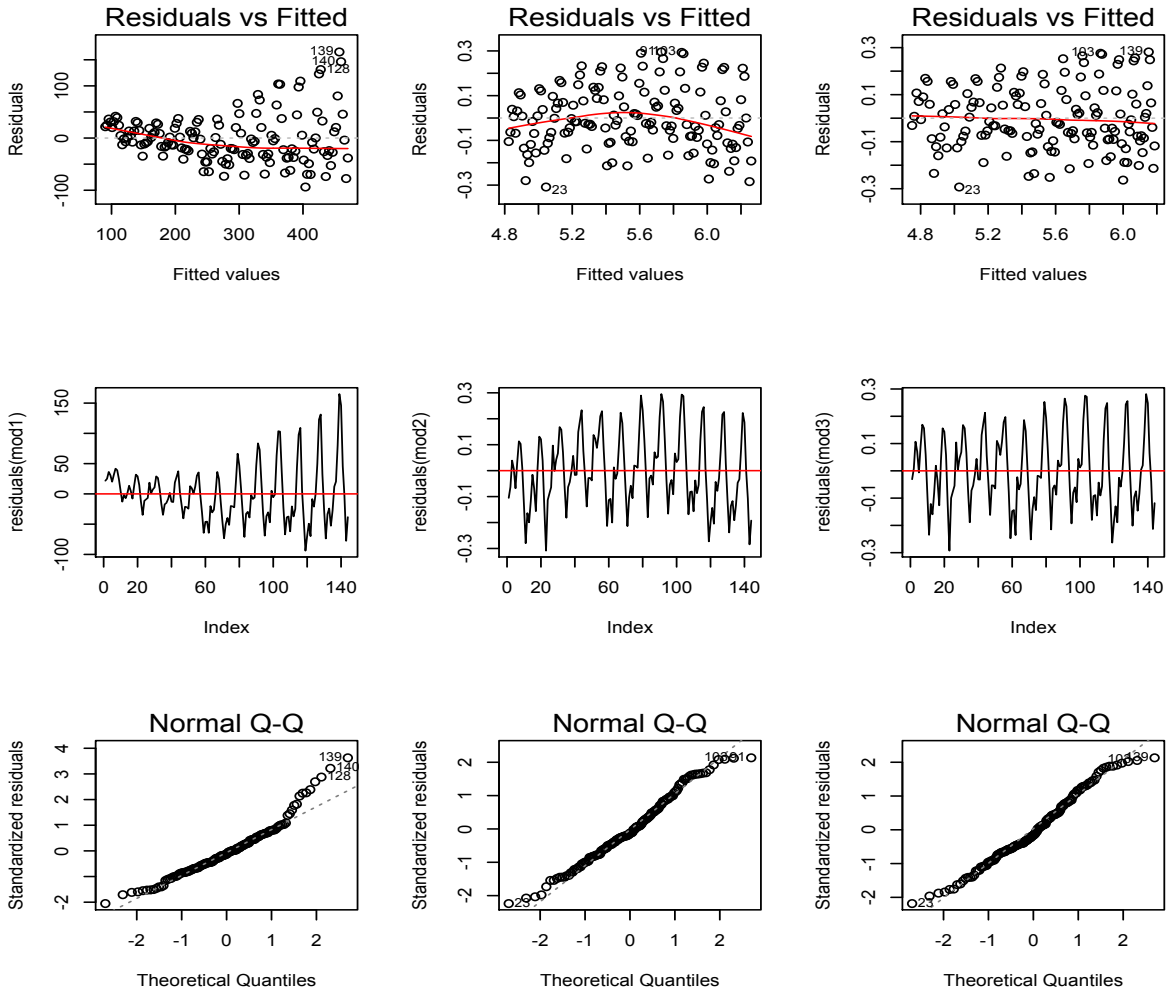


Figure 3.1 – Étude de résidus pour les trois modèles

Pour comprendre la saisonnalité, nous examinons le month plot du trafic. La figure 3.2 montre un diagramme séquentiel pour chacune des saisons supposées de la série. S'il n'y avant pas d'effet saisonnier, ces 12 chronogrammes se ressembleraient. On note que les chronogrammes sont presque droits et parallèles. Juillet montre la plus forte activité.

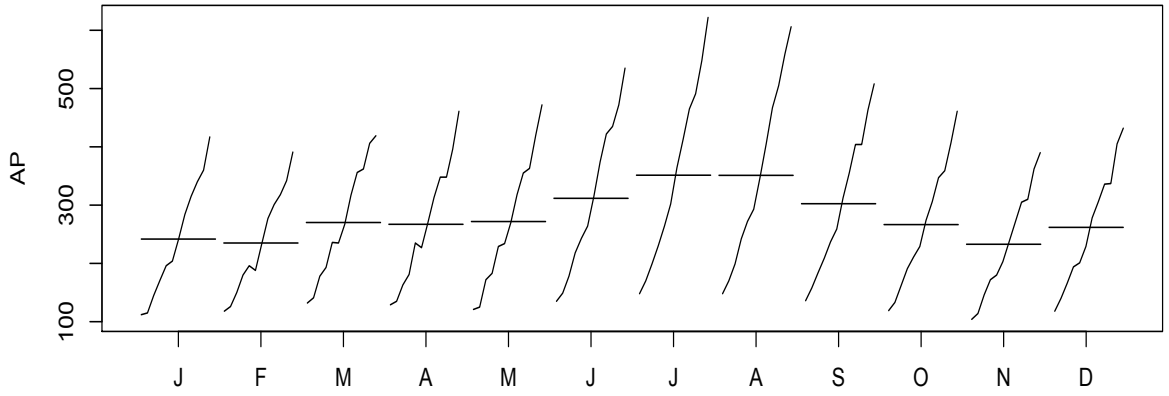


Figure 3.2 – Month plot du trafic

Janvier, février et novembre sont les mois de plus faible activité. D'après le month plot, on propose le modèle suivant,

$$\log Y_{i,j} = \mu_j + \beta_1 t_{i,j} + \beta_2 (t_{i,j} - \bar{t})^2 + \varepsilon_{i,j}, \quad i : \text{année}, \quad j : \text{mois}.$$

La forme matricielle du modèle est la suivante

$$\log \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ \vdots \\ Y_{1,12} \\ \vdots \\ Y_{12,1} \\ Y_{12,2} \\ \vdots \\ Y_{12,12} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_{12} \end{pmatrix} + t\beta_1 + (t - \bar{t})^2 \beta_2 + \varepsilon.$$

Chapitre 4

Analyse de la variance

4.1 Un exemple simple

Un forestier s'intéresse aux hauteurs moyennes de trois forêts. Pour les estimer, il échantillonne un certain nombre d'arbres et mesure leurs hauteurs.

Forêt 1	$n_1 = 6$	23,4	24,4	24,6	24,9	25,0	26,2
Forêt 2	$n_2 = 5$	22,5	22,9	23,7	24,0	24,4	
Forêt 3	$n_3 = 7$	18,9	21,1	21,2	22,1	22,5	23,5	24,5

Ces données peuvent être présentées de deux minières équivalentes.

1. On dispose de trois échantillons indépendants et on désire comparer leurs moyennes.

2. On dispose d'un seul échantillon de longueur 18 et d'une variable explicative qualitative, ou facteur, le numéro de la forêt. En prenant ce second point de vue, on parle d'analyse de la variance à un facteur.

4.1.1 Modèle statistique

Soit Y_{ij} la hauteur du $j^{ème}$ arbre de la forêt i, soit μ_i la hauteur moyenne de la forêt i. Dans ce cadre, un modèle possible est le suivant,

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \tag{1}$$

où ε_{ij} est la variabilité de l'arbre j par rapport à la hauteur moyenne de la forêt i.

Si les forêts sont équivalentes, c'est-à-dire la moyenne de la hauteur des arbres est la même dans chaque forêt, on peut proposer le sous-modèle suivant,

$$Y_{ij} = \mu + \varepsilon_{ij},$$

où μ est la valeur commune de μ_1, μ_2 et μ_3.

Comme précédemment, nous allons faire quelques hypothèses sur les ε_{ij}, pour $i = 1, \ldots, I$ et $j = 1, \ldots, n_i$,

$\mathbb{E}(\varepsilon_{ij}) = 0$

$\text{Var}(\varepsilon_{ij}) = \sigma^2$

ε_{ij} sont i.i.d. de loi gaussienne.

Dans le premier modèle on déduit que la valeur prédite de Y_{ij} est

$$\hat{Y}_{ij} = Y_{i.} = \hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij},$$

où n_i est le nombre d'arbre de la forêt i.

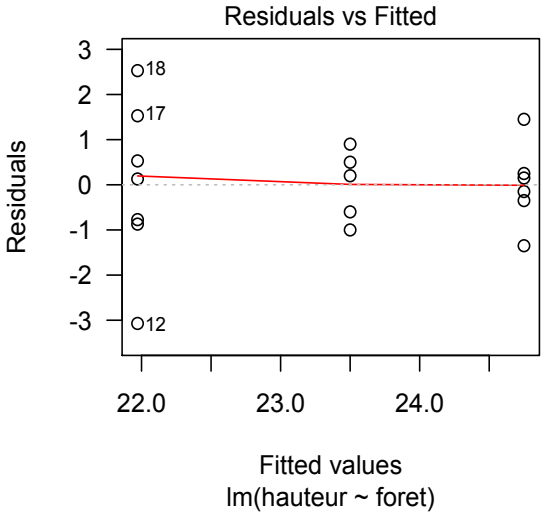
Dans le sous-modèle on déduit que la valeur prédite de Y_{ij} est

$$\hat{Y}_{ij} = Y_{..} = \hat{\mu} = \frac{1}{n_1 + n_2 + n_3} \sum_{i,j} Y_{ij}.$$

Table d'analyse de la variance

Source	Somme de carrés	Degré de liberté	Carré moyen
inter-groupe	$\sum_{i,j}(Y_{i.} - Y_{..})^2$	$I - 1$	$\frac{1}{I-1}\sum_{i,j}(Y_{i.} - Y_{..})^2 := \text{SCM}_{\text{inter}}$
intra-groupe	$\sum_{i,j}(Y_{ij} - Y_{i.})^2$	$n - I$	$\frac{1}{n-I}\sum_{i,j}(Y_{ij} - Y_{i.})^2 := \text{SCM}_{\text{intra}}$
totale	$\sum_{i,j}(Y_{ij} - Y_{..})^2$	$n - 1$	$\frac{1}{n-1}\sum_{i,j}(Y_{ij} - Y_{..})^2$

D'après un calcul simple on obtient que les moyennes de la hauteur des arbres de trois forêts sont respectivement 24,75 23,5 et 21,97. Le graphique des résidus en fonction des valeurs prédites est présenté ci dessous. Les numéros 12, 17 et 18 sont les numéros des trois individus ayant les valeurs absolues de résidus maximales.



4.1.2 Test de Fisher

On va tester l'hypothèse suivante.

H_0 : le modèle vérifie $\mu_1 = \mu_2 = \mu_3$ contre H_1 : le modèle ne vérifie pas $\mu_1 = \mu_2 = \mu_3$.

La règle de décision s'écrit

$$H_0 \Leftrightarrow \hat{F} = \frac{\text{SCM}_{\text{inter}}}{\text{SCM}_{\text{intra}}} \leq F_{(I-1,n-I),1-\alpha} \quad \text{avec} \quad \alpha = \mathbb{P}\left(F_{(I-1,n-I)} > F_{(I-1,n-I),1-\alpha}\right).$$

4.1.3 Test de Student

Si on choisit deux populations à comparer, par exemple la 1 et la 2, on déduire un test de Student de comparaison de ces deux moyennes. L'hypothèse sur laquelle on va tester est la suivante.

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2.$$

La statistique de test $\hat{T}$ est définie par

$$\hat{T} = |Y_{1\cdot} - Y_{2\cdot}| \left(\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)^{-1/2}$$

où $\hat{\sigma}^2 = \frac{1}{n-I} \sum_{i,j} (Y_{i\cdot} - Y_{\cdot\cdot})^2$. Pour rejeter l'hypothèse H_0, il faudra que $\hat{T} > T_{n-I,1-\alpha/2}$ où $T_{n-I,1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-I$ degrés de liberté.

Sortie d'exemple traité par R

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.7500     0.5419  45.672  < 2e-16 ***
foret2       -1.2500     0.8038  -1.555  0.14075
foret3       -2.7786     0.7385  -3.763  0.00188 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.327 on 15 degrees of freedom
Multiple R-squared: 0.4877,Adjusted R-squared: 0.4194
F-statistic:  7.14 on 2 and 15 DF,  p-value: 0.006627
```

4.1.4 Utilisation de variables muettes

La variable muette (aussi connue comme la variable dummy/binaire/d'indicateur) est une variable qui ne prend que les valeurs 0 ou 1 pour indiquer la présence ou l'absence de certains effets catégoriques qui ne peuvent être attendus pour déplacer le résultat. Supposons que la taille d'échantillon est 2 pour chaque forêt. Nous avons

Forêt 1	$n_1 = 2$	23,4	24,4
Forêt 2	$n_2 = 2$	22,5	22,9
Forêt 3	$n_3 = 2$	18,9	21,1

La forme matricielle du modèle (1) est la suivante

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}.$$

En fait, cela est le modèle de détermination de Y par trois variables explicatives qualitatives D_1, D_2, et D_3,

$$Y_i = D_1\mu_1 + D_2\mu_2 + D_3\mu_3 + \varepsilon_i, \quad i = 1, \ldots, 6,$$

où

$$D_j = \begin{cases} 1 & \text{si l'observation } Y_i \text{ appartient à la forêt } j, \\ 0 & \text{sinon,} \end{cases} \quad j = 1, 2, 3.$$

4.2 Analyse de la variance à deux facteurs croisés

On compare l'action de deux traitements désinfectants sur des échantillons de racines de dents de vaches contaminées au préalable par deux sources de germes. La réponse est le logarithme du nombre moyen de germes restants. Pour chaque combinaison de germe et traitement nous avons trois réponses. Le tableau des observations est le suivant.

germe1 : trait1	$n_{11} = 3$	2,00	0,94	0,12
germe1 : trait2	$n_{12} = 3$	-0,18	0,65	0,40
germe2 : trait1	$n_{21} = 3$	-0,40	0,94	1,84
germe2 : trait2	$n_{22} = 3$	0,38	-0,14	0,80

Le modèle peut s'écrire

$$Y_{ijk} = \theta_{ij} + \varepsilon_{ijk}, \quad i = 1, 2, j = 1, 2, k = 1, 2, 3.$$

Pour l'instant, le modèle posé est en fait un modèle d'analyse de la variance à un facteur : le facteur produit « germe : trait ». Pour apparaître l'interaction entre deux facteurs, on définit les paramètres suivants,

$\mu = \theta_{..}$: la moyenne générale,

$\alpha_i = \theta_{i.} - \theta_{..}$: l'effet différentiel de la modalité i du premier facteur,

$\beta_j = \theta_{.j} - \theta_{..}$: l'effet différentiel de la modalité j du deuxième facteur,

$\gamma_{ij} = \theta_{ij} - \theta_{i.} - \theta_{.j} + \theta_{..}$: l'effet d'interaction entre la modalité i du 1er facteur et la modalité j du 2ème facteur.

Au final, le modèle précédant se réécrit sous la forme

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}.$$

Remarque 14. Lorsque les paramètres d'interaction γ_{ij} sont nuls pour tous $(i, j) \in \{1, \ldots, I\} \times \{1, \ldots, J\}$, le modèle est dit additif. Sinon, on dit que le modèle est avec interaction.

On parle d'« effets principaux » pour tout ce qui est relatif aux paramètres α_i et β_j. On veut tester les différentes hypothèses stipulant la présence ou non d'un des effets principaux ou de l'interaction. Plus précisément on définit les hypothèses suivantes,

$H_0^{(1)}$: tous les coefficients α_i sont nuls,

$H_0^{(2)}$: tous les coefficients β_j sont nuls,

$H_0^{(3)}$: tous les coefficients γ_{ij} sont nuls.

Les statistiques $\hat{F}$ présentées dans la table suivante sont utilisées pour tester ces hypothèses.

Table d'analyse de la variance

Source	Somme de carrés	Degré de liberté	$\hat{F}$
Facteur 1	$\sum_{i,j,k}(Y_{i..} - Y_{...})^2 := \text{SCE1}$	$I-1$	$\frac{(n-I\cdot J)}{(I-1)}\frac{\text{SCE1}}{\text{SCR}}$
Facteur 2	$\sum_{i,j,k}(Y_{.j.} - Y_{...})^2 := \text{SCE2}$	$J-1$	$\frac{(n-I\cdot J)}{(J-1)}\frac{\text{SCE2}}{\text{SCR}}$
Interaction	$\sum_{i,j,k}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2 := \text{SCI}$	$(I-1)(J-1)$	$\frac{(n-I\cdot J)}{(I-1)(J-1)}\frac{\text{SCI}}{\text{SCR}}$
Résiduelle	$\sum_{i,j,k}(Y_{ijk} - Y_{ij.})^2 := \text{SCR}$	$n-I\cdot J$	

Bibliographie

[1] Yves Aragon, *Séries temporelles avec R*. Springer, 2011.

[2] Jean-Marc Azaïs et Jean-Marc Bardet, *Le modèle linéaire par l'exemple : Régression, Analyse de la variance et Plans d'expérience illustrés avec R, SAS et Splus*. Dunod, 2006.

[3] Joseph Rynkiewicz, *Statistiques et Probabilités L2*. polycopié du cours "Statistiques et Probabilités" en MASS L2 2011.

[4] Gilles Fäy, *Probabilités et Statistiques*. polycopié du cours "Probabilités et Statistiques" en Licence Aménagée Parcours SVTE 2009-2010.

[5] Virginie Delsart, Arnaud Rys et Nicolas Vaneecloo, *Économétrie théorie et application sous SAS*. Septentrion, 2009.