Sessió 12

Introducció a les sèries temporals amb R

12.1 Introducció

Una *sèrie temporal* és una successió d'observacions ordenades segons una seqüència temporal.

Exemples:

- Producció d'automòbils d'un país en diferents anys.
- Consum d'electricitat (mitjana familiar) en diferents mesos de diversos anys.
- Població activa d'un país en diferents mesos.
- Temperatura de Bellaterra (mitjana diària) al llarg dels dies.
- Índex (IBEX 35) al mercat continu (per exemple, cada segon).
- Ingressos hospitalaris a Catalunya, per dies.

Estudiem conjuntament dues variables:

1. Mesura del temps: T discretitzada (valors $t_1, \ldots, t_n$ o bé $1, \ldots, n$).

   Per exemple, si T es mesura en anys, $t : 2000, 2001, \ldots, 2019$. O bé si es mesura en dies, T podria variar del 1 d'octubre de 2019 al 10 de desembre de 2019, ...

2. Variable quantitativa: Y (valors $y_1, \ldots, y_n$)

Són dades emparellades, de la forma

T	Y
t_1	y_1
$\vdots$	$\vdots$
t_n	y_n

o bé

T	Y
1	y_1
$\vdots$	$\vdots$
n	y_n

El que és important és que tenim una sola observació de Y per a cada temps $T = t_i$ (o cada $T = i$).

Podem, d'entrada, fer un estudi i representació gràfica bivariant (models de regressió, etc.)

També podem avaluar índexs consecutius: globals, mitjans, etc.

Finalment, com veurem, es poden estudiar les sèries temporals afegint al model bivariant un altre ingredient (l'anomenada *component estacional*).

12.2 Exemple

Començarem amb un exemple, treballant amb R. Es tracta de la sèrie de naixements a Nova York els anys 1947–1959 (en milers).

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1947	21.44	21.09	23.71	21.67	21.75	20.76	23.48	23.82	23.11	23.11	21.76	22.07
1948	21.94	20.04	23.59	21.67	22.22	22.12	23.95	23.50	22.24	23.14	21.06	21.57
1949	21.55	20.00	22.42	20.61	21.76	22.87	24.10	23.75	23.26	22.91	21.52	22.02
1950	22.60	20.89	24.68	23.67	25.32	23.58	24.67	24.45	24.12	24.25	22.08	22.99
1951	23.29	23.05	25.08	24.04	24.43	24.67	26.45	25.62	25.01	25.11	22.96	23.98
1952	23.80	22.27	24.77	22.65	23.99	24.74	26.28	25.82	25.21	25.20	23.16	24.71
1953	24.36	22.64	25.57	24.06	25.43	24.64	27.01	26.61	26.27	26.46	25.25	25.18
1954	24.66	23.30	26.98	26.20	27.21	26.12	26.71	26.88	26.15	26.38	24.71	25.69
1955	24.99	24.24	26.72	23.48	24.77	26.22	28.36	28.60	27.91	27.78	25.69	26.88
1956	26.22	24.22	27.91	26.98	28.53	27.14	28.98	28.17	28.06	29.14	26.29	26.99
1957	26.59	24.85	27.54	26.90	28.88	27.39	28.07	28.14	29.05	28.48	26.63	27.73
1958	27.13	24.92	28.96	26.59	27.93	28.01	29.23	28.76	28.41	27.95	25.91	26.62
1959	26.08	25.29	27.66	25.95	26.40	25.57	28.86	30.00	29.26	29.01	26.99	27.90

```
births <- scan("http://robjhyndman.com/tsdldata/data/nybirths.dat")
class(births)    # és un vector de les observacions
naixements<-births[-(1:12)]  # treiem els primers casos
length(naixements)
[1] 156
temps<-1:156
plot(temps,naixements,type="l",xlim=c(1,156),col=2,
      main="Naixements a la ciutat de Nova York",axes=F)
axis(1,at=seq(1,157,by=12),labels=as.character(seq(1947,1960)))
axis(2)
abline(v=seq(1,157,by=12),col="lightgray")
```

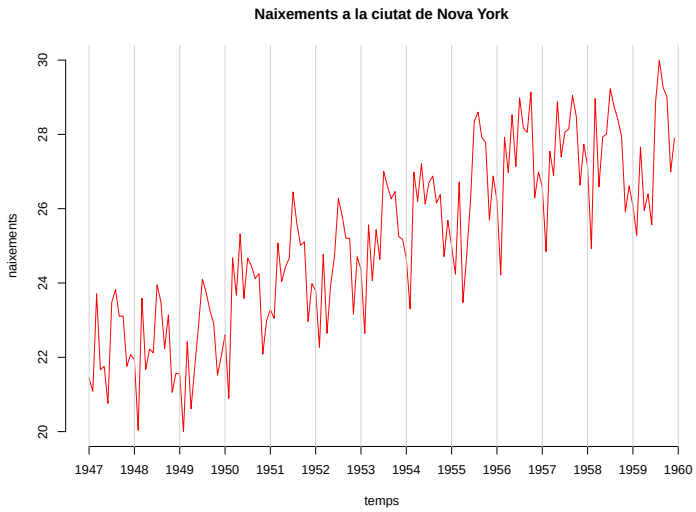


Figura 1: Representació de la sèrie temporal, amb els períodes marcats en gris clar

Mostrarem el procediment en tres etapes:

1. Calculem la **recta de regressió de** Y **sobre** T i la dibuixem al plot anterior

```
model<-lm(naixements~temps)
a<-model$coefficients[1]
b<-model$coefficients[2]
abline(a,b,col=4)
```

La recta de regressió és $y = 21.684 + 0.0438\,t$, que s'anomena **tendència de la sèrie**. Fixem-nos en el gràfic, alguns mesos el nombre de naixements està per sobre de la tendència i en altres està per sota.

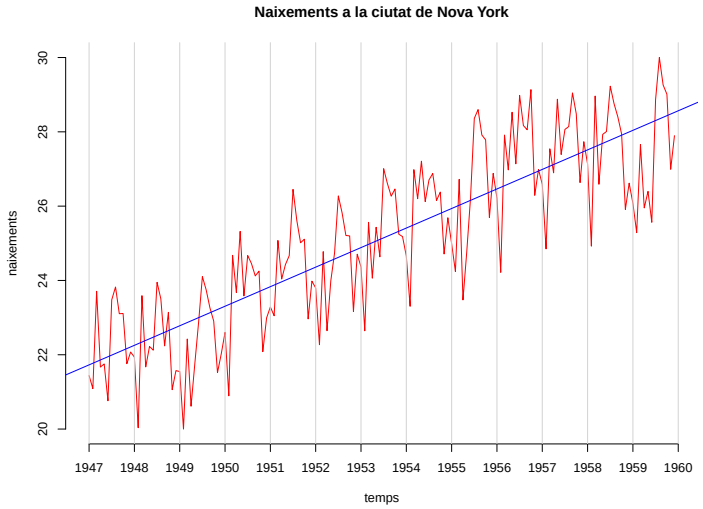


Figura 2: Representació de la sèrie temporal i la tendència (la recta de regressió)

2. Ara volem fer una estimació de les fluctuacions que fa la sèrie al voltant de la tendència. Seran números e_{mes}, un número per a cada mes (hem de calcular $e_1, \ldots, e_{12}$). S'anomenen **coeficients estacionals**.

   Quan tinguem els coeficients estacionals calculats, l'estimació (aproximació) de la sèrie serà

$$Y \approx \underbrace{a + bT}_{\text{tendència}} + \underbrace{e_{\text{mes}}}_{\text{coef.estacional del mes}} \iff Y - (a + bT) \approx e_{\text{mes}}$$

   Mètode per calcular els coeficients estacionals:

   2.a Fer la diferència entre la sèrie i la tendència. Aleshores les fluctuacions seran al voltant del zero.

```
SenseTend<-naixements-(a+b*temps)
plot(temps,SenseTend,type="l",xlim=c(1,156),col=2,axes=F,
     main="diferència entre la sèrie i la tendència")
axis(1,at=seq(1,157,by=12),labels=as.character(seq(1947,1960)))
axis(2)
abline(v=seq(1,157,by=12),col="lightgray")
abline(h=0,col=4)
```

   2.b e_1 és la mitjana de `SenseTend` per als mesos de gener (és a dir, per a $t = 1, 13, \ldots$), e_2 és la mitjana de `SenseTend` per als mesos de febrer (per a $t = 2, 14, \ldots$.

   La millor manera de calcular-los és col·locar en una matriu el vector `SenseTend` tal com estan les dades a la primera taula (fixem-nos que hem de llegir per files):

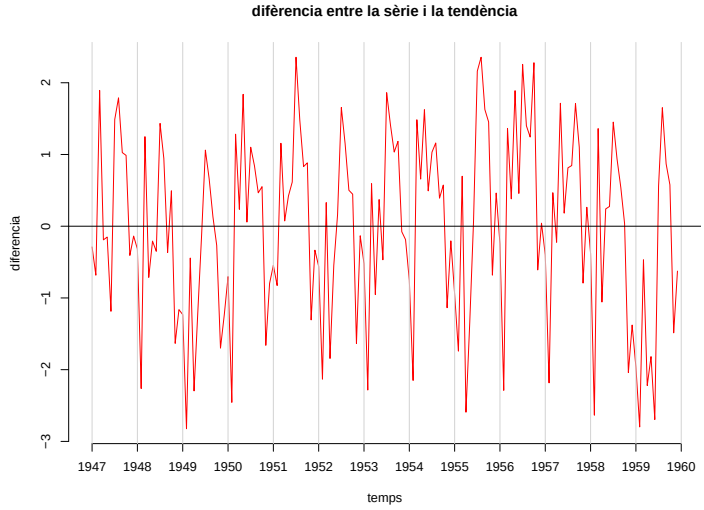


Figura 3: Representació de la nova sèrie temporal SenseTend (diferència entre Y i la tendència)

```
matriudif<-matrix(SenseTend,ncol=12,byrow=T)
dimnames(matriudif)<-list(1947:1959, month.abb)
round(matriudif,2)
```

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1947	-0.29	-0.68	1.89	-0.19	-0.15	-1.19	1.49	1.79	1.03	0.99	-0.41	-0.14
1948	-0.32	-2.26	1.25	-0.71	-0.21	-0.35	1.43	0.94	-0.37	0.49	-1.63	-1.16
1949	-1.23	-2.82	-0.44	-2.30	-1.19	-0.13	1.06	0.66	0.13	-0.27	-1.70	-1.24
1950	-0.70	-2.46	1.28	0.24	1.84	0.06	1.10	0.84	0.47	0.55	-1.66	-0.80
1951	-0.55	-0.83	1.16	0.07	0.42	0.62	2.36	1.48	0.83	0.88	-1.31	-0.33
1952	-0.56	-2.13	0.33	-1.84	-0.55	0.16	1.65	1.15	0.50	0.45	-1.63	-0.13
1953	-0.52	-2.28	0.59	-0.95	0.37	-0.47	1.86	1.41	1.03	1.18	-0.08	-0.19
1954	-0.75	-2.15	1.48	0.66	1.62	0.49	1.03	1.16	0.39	0.57	-1.14	-0.20
1955	-0.95	-1.74	0.70	-2.59	-1.34	0.06	2.16	2.36	1.63	1.45	-0.68	0.46
1956	-0.25	-2.29	1.36	0.38	1.89	0.46	2.26	1.40	1.24	2.28	-0.61	0.04
1957	-0.40	-2.18	0.47	-0.22	1.71	0.18	0.81	0.85	1.71	1.10	-0.79	0.26
1958	-0.38	-2.63	1.36	-1.06	0.24	0.28	1.45	0.94	0.54	0.04	-2.04	-1.38
1959	-1.96	-2.80	-0.47	-2.22	-1.82	-2.69	0.56	1.65	0.87	0.58	-1.49	-0.63

Fixem-nos que totes les diferències corresponents al mes de gener són negatives, o totes les de l'agost positives. Ja podem calcular els coeficients estacionals:

```
est<-apply(matriudif,2,mean)
est
```

Els coeficients estacionals són

$$\begin{array}{llll} e_1 = -0.681 & e_2 = -2.097 & e_3 = 0.843 & e_4 = -0.827 \\ e_5 = 0.219 & e_6 = -0.194 & e_7 = 1.480 & e_8 = 1.279 \\ e_9 = 0.769 & e_{10} = 0.792 & e_{11} = -1.167 & e_{12} = -0.417 \end{array}$$

3. Ja tenim l'**aproximació de la sèrie** de naixements. D'una banda la tendència (recta de regressió) i de l'altra els coeficients estacionals, dels quals n'hi ha dotze i que es van repetint periòdicament. Fem una representació gràfica de la sèrie i de l'aproximació:

```
component.estacional<-rep(est,13)
tendencia<-a+b*temps
estimacio<- tendencia + component.estacional
plot(temps,naixements,type="l",xlim=c(1,156),col=1,
    main="Naixements a la ciutat de Nova York", axes=F)
axis(1,at=seq(1,157,by=12),labels=as.character(seq(1947,1960)))
axis(2)
points(temps,estimacio,type="l",col=2)
legend(1,30,legend=c("sèrie naixements", "aproximació"),
       col=c("black", "red"), lty=1, cex=0.9)
```

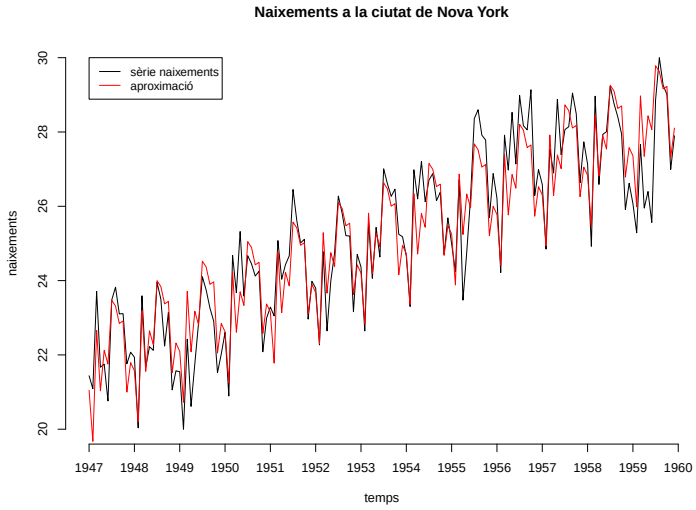


Figura 4: Representació de la sèrie temporal i la seva aproximació (tendència més estacionalitat)

4. Podem utilitzar l'aproximació de la sèrie per **fer prediccions:**

   Suposem que volem conèixer el nombre de naixements el febrer de 1963, que està més enllà del rang temporal de la sèrie. Si $t = 156$ és el darrer període de la sèrie i correspon a desembre de 1959, cal sumar-hi 3 anys sencers (1960, 1961 i 1962) i 2 mesos (gener i febrer) serà $t = 156 + 3 * 12 + 2 = 194$. El coeficient estacional que necessitem és $e_2 = -2.097$ (el coeficient estacional de febrer):

```
t1<-156+3*12+2
yestim<-a+b*t1+est[2]
round(yestim,3)
28.092
```

$$\hat{y} = a + bt + e_2 = 21.684 + 0.0438 \cdot 194 - 2.097 = 28.092.$$

12.3 Model d'estimació d'una sèrie

En una sèrie s'aprecien diferents components:

1. la tendència
2. les variacions estacionals
3. les fluctuacions cícliques (no les treballarem)
4. el component residual (no se'n fa estimació, és la diferència entre la sèrie i el model que ajustem)

L'anàlisi d'una sèrie temporal consisteix en:

- Descripció de pautes de regularitat dels seus components, en l'interval de temps del qual disposem les dades.
- Predicció de la seva evolució futura, suposant que les pautes es mantenen en el futur.

La tendència

És el comportament mitjà de la sèrie a llarg termini.

Tendència lineal: en algunes sèries, al fer la representació de la variable Y en funció del temps i calcular la recta de regressió (de Y sobre T) veiem que hi ha un bon ajust lineal. En la representació gràfica s'observa que la sèrie fluctua al voltant de la recta, amb oscil·lacions que es repeteixen periòdicament. En aquest cas **la tendència és la recta de regressió**.

Tendència exponencial: en algunes sèries gràficament s'observa que la tendència és més aviat exponencial i que la fluctuació de la sèrie al voltant de la tendència es va amplificant amb el temps.

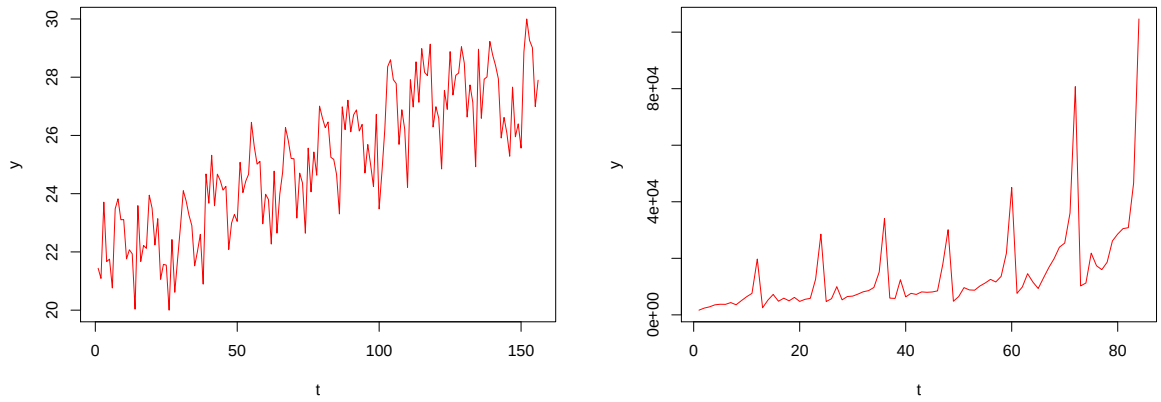


Figura 5: A l'esquerra, sèrie amb tendència lineal i a la dreta, sèrie amb tendència exponencial.

La component estacional (quan la tendència és lineal)

Ara restem a la sèrie (Y) la tendència ($a + bT$), obtenint una nova sèrie, que anomenem `SenseTend`. Al representar-la veiem que oscil·larà al voltant del zero (fixeu-vos amb la Figura 3). La sèrie `SenseTend` conté la component estacional, la component cíclica (que no considerem) i la component residual de la sèrie original Y.

Per a cada estació, calcularem un número, anomenat coeficient estacional. Si per exemple, treballem per estacions, tindrem 4 coeficients estacionals.

El coeficient de la primavera s'obté com la mitjana de `SenseTend` només per als valors de la sèrie corresponents a primavera, el coeficient estacional de l'estiu el trobarem fent la mitjana dels valors de Y corresponents a l'estiu, etc.

La component estacional serà anar repetint la successió dels 4 coeficients estacionals.

El model additiu

Un cop hem calculat la tendència i els coeficients estacionals, tenim una aproximació de la sèrie que ens permetrà fer predicions:

$$\boxed{Y \approx (a + bT) + e} \tag{12.1}$$

on $y = a + bt$ és la recta de regressió i e és la successió periòdica que conté els coeficients estacionals. Aquest model s'aplica a una sèrie com la de la Figura 5, esquerra.

El model multiplicatiu

Si detectem que la tendència és exponencial i que s'amplifiquen les oscil·lacions de la sèrie, com a la Figura 5 dreta, transformem Y aplicant-li el logaritme.

Ara $\log(Y)$ és una nova sèrie temporal que al representar-la oscil·larà al voltant d'una recta sense amplificar oscil·lacions, com a la Figura 5, esquerra.

Ajustem el model additiu (12.1) a la sèrie $\log(Y)$:

$$\log(Y) \approx a + bT + e_t,$$

on $a + bt$ és la recta de regressió de $\log(Y)$ sobre T i e_t la successió de coeficients estacionals de la sèrie $\log(Y)$. Per tant

$$\boxed{Y \approx e^{a+bT+e_t} = e^{e_t}\, e^{a+bt} = \tilde{e}_t\, e^{a+bt}}$$

és l'aproximació de Y que buscàvem (on hem posat $\tilde{e_t} = e^{e_t}$). Fixem-nos que les variacions estacionals $\tilde{e_t}$ estan multiplicades per la tendència e^{a+bt} de manera que quant més gran sigui la tendència més s'amplifica l'oscil·lació al voltant de la tendència. En el cas lineal, en canvi, els coeficients estacionals e_t se sumaven a la tendència i per tant sempre eren oscil·lacions igual d'àmplies.

Prediccions

Utilitzem les fórmules obtingudes, buscant quina és la t que correspondria a la predicció i tenint en compte en quina "estació" estem (estació o mes, o hora,...) i quin coeficient hem d'utlitzar.

12.4 Exercicis de pràctica

1. Considerem les vendes d'una empresa de begudes carbòniques en funció de les 4 estacions de l'any:

	any 1	any 2	any 3	any 4
primavera	2.0	2.2	2.2	2.4
estiu	3.1	3.0	3.5	3.6
tardor	2.6	2.8	4.3	4.5
hivern	1.8	2.0	2.1	2.2

   (a) Obtingueu, mitjançant regressió lineal, la funció de tendència lineal, $a + bT$. Representeu-la gràficament conjuntament amb la sèrie inicial.

   (b) Calculeu les variacions estacionals de la sèrie, e_i.

   (c) Calculeu la sèrie de prediccions $\hat{Y} = a + bT + e$. Representeu-la gràficament conjuntament amb la sèrie inicial.

   (d) Feu la predicció de vendes per a la primavera de l'any 5.

2. Un institut de Secundària decideix estudiar l'evolució de la mitjana de les notes de matemàtiques dels seus estudiants. Per tal de fer això, anota després de cada avaluació (se suposa que a cada curs n'hi ha 3) la mitjana que han obtingut els alumnes. Després de 5 anys s'obté la taula següent

	any 1	any 2	any 3	any 4	any 5
1a avaluació	4.5	4.8	5	5.1	5.2
2a avaluació	4.6	4.9	5	5.2	5.3
3a avaluació	4.4	4.7	4.9	5.1	5.1

Considerem les dades com una sèrie temporal $Y_t =$ "nota en el període t", amb $t = 1, 2, \ldots, 15$ ($t = 1$ és la primera avaluació de l'any 1 i $t = 15$ és la 3a avaluació de l'any 5).

(a) Calculeu la recta de regressió de Y sobre T (la tendència).

(b) Representeu conjuntament la sèrie i la tendència. Què observeu?

(c) Calculeu els coeficients estacionals de la sèrie. Interpreteu-los.

(d) Quina seria la nota mitjana prevista en la tercera avaluació de l'any 6?

3. Les dades següents corresponen nombre de passatgers mensuals d'una aerolínia, en el període 1949-1960 (les trobareu al dataset `AirPassengers` de la llibreria `datasets` de R, que es carrega automàticament).

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	317	313	318	374	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	606	508	461	390	432

(a) Representeu gràficament la sèrie de les dades i comproveu que les variacions estacionals es van amplificant, i la tendència no sembla que permeti un ajust lineal. Calculeu la recta de regressió i representeu-la conjuntament amb la sèrie.

(b) Transformeu la sèrie Y_t logarítimicament i representeu la nova sèrie $Z_t = \log(Y_t)$.

(c) Calculeu la tendència de la sèrie $Z_t = \log(Y_t)$ (amb un model additiu amb tendència $a + bT$). I representeu-la juntament amb la sèrie Z_t.

(d) De l'apartat anterior, deduïu la tendència per a Y_t serà e^{a+bt} (ajustant d'aquesta manera un model multiplicatiu a Y_t).

(e) Representeu gràficament Y_t i la seva tendència obtinguda a l'apartat anterior i compareu amb la representació de l'aparatat (a).

(f) Calcueu els coeficients estacionals de $Z_t = \log(Y_t)$ i després els coeficients corresponents del model multiplicatiu per a Y_t.

(g) Calcueu $\hat{y}_t = e^{e_i}\, e^{bt+a} = c_i\, e^{bt+a}$, on $y = bt + a$ és la recta de regressió de $Z_t = \log(Y_t)$ sobre T de l'apartat (c). Afegiu $\hat{Y}_t$ al gràfic de l'apartat (e).

(h) Feu una predicció per al nombre de passatgers l'abril de 1961.

4. Les següents dades corresponen a vendes de souvenirs en una botiga d'un poble turístic de Queensland, Austràlia de Gener de 1987 a Desembre de 1993 (dades de Wheelwright and Hyndman, 1998)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1987	1665	2398	2841	3547	3753	3715	4350	3566	5022	6423	7601	19756
1988	2500	5198	7225	4806	5901	4951	6179	4752	5496	5835	12600	28542
1989	4717	5703	9958	5305	6492	6631	7350	8177	8573	9690	15152	34061
1990	5921	5815	12421	6370	7609	7225	8121	7979	8093	8477	17915	30114
1991	4827	6470	9639	8821	8722	10209	11277	12552	11637	13607	21822	45061
1992	7615	9850	14558	11587	9333	13082	16733	19889	23933	25391	36025	80722
1993	10243	11267	21827	17357	15998	18602	26155	28587	30505	30821	46634	104661

(a) Escanegeu les dades amb

```
souvenir <- scan("http://robjhyndman.com/tsdldata/data/fancy.dat")
```

(b) Representeu gràficament la sèrie. Observeu que les oscil·lacions temporals es van amplificant amb el temps.

(c) Calculeu la recta de regressió, representeu-la amb la sèrie i comproveu que l'ajust lineal és molt dolent, amb la recta de regressió i el coeficient de determinació.

(d) Apliqueu logaritmes a la sèrie i feu la representació gràfica de $\log(Y)$ respecte de T. Observeu que l'ajust és millor.

(e) Treballant com en el problema 3 trobeu primer la tendència $a+bT$ i els coeficients estacionals de la sèrie $\log(Y)$.

(f) Representeu conjuntament $\log(Y)$ i l'estimació $a + bT + e_i$.

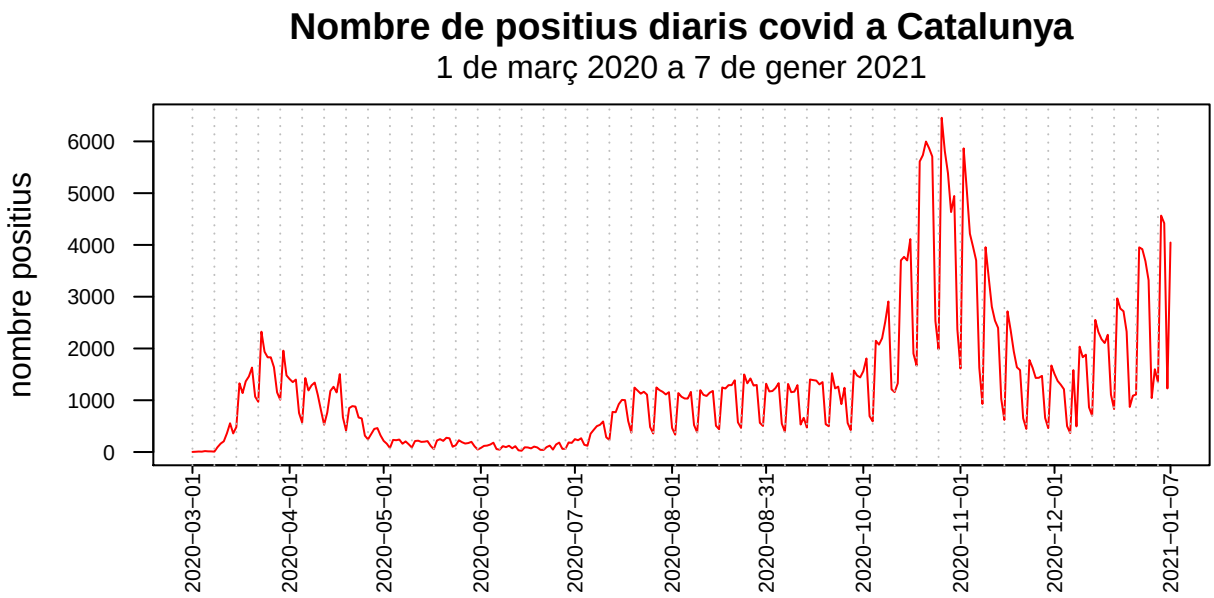
(g) Representeu conjuntament Y i l'estimació e^{a+bT+e_i}

(h) Feu una previsió de les vendes per al maig del 1994.

5. Al Moodle hi trobareu el fitxer de dades de R, `covid.RData` que conté els vectors `incidencia` i `dates`, que corresponen a la incidència diària COVID a Catalunya en les dates del vector de caràcters `dates`.

   El fitxer s'ha extret de `https://dadescovid.cat/descarregues` (totals població total, Catalunya). Però s'han hagut de sumar abans tots els casos que corresponen al mateix dia. Per si més endavant voleu treballar amb actualitzacions de les dades, incloem el codi per passar de les dades descarregades als dos vectors que us donem al fitxer `covid.RData`. També incloem el codi per fer una representació gràfica

```
covid<-read.csv2("catalunya_diari_total_pob.csv")
covid$DATA<-as.factor(covid$DATA)
incidencia<- as.vector(by(covid$CASOS_CONFIRMAT,covid$DATA,sum))
dates<-levels(covid$DATA)
n<-length(incidencia)
plot(1:n,incidencia,type="l",ylab="nombre positius",xlab="",xaxt="n",
    col=2,las=2,main="Nombre de positius diaris covid a Catalunya",
    cex=0.7,cex.axis=0.7)
axis(side=1,c(1,32,62,93,123,154,184,215,246,276,313),
    tcl=-0.2,labels=FALSE)
mtext(side=3,line=0.5,"1 de març 2020 a 7 de gener 2021")
mtext(at=c(1,32,62,93,123,154,184,215,246,276,313),
      text=dates[c(1,32,62,93,123,154,184,215,246,276,313)],
      side=1,cex=0.7,las=2,line=0.2)
abline(v=seq(1,309,by=7),col=8,lty=3)
```



L'exercici consisteix en triar un interval de dades que us sembli adequat (que s'hi pugui ajustar una tendència lineal o bé exponencial), per exemple del 2020-09-15 al 2020-11-01, o del 2020-12-01 a l'actualitat. I ajusteu un model additiu i un multiplicatiu, compareu-los. És important que us agafeu un nombre sencer de setmanes. Dels dos models, el que creieu que s'ajusta millor (en vista dels gràfics), feu una predicció per a 10 dies més que la darrera data.