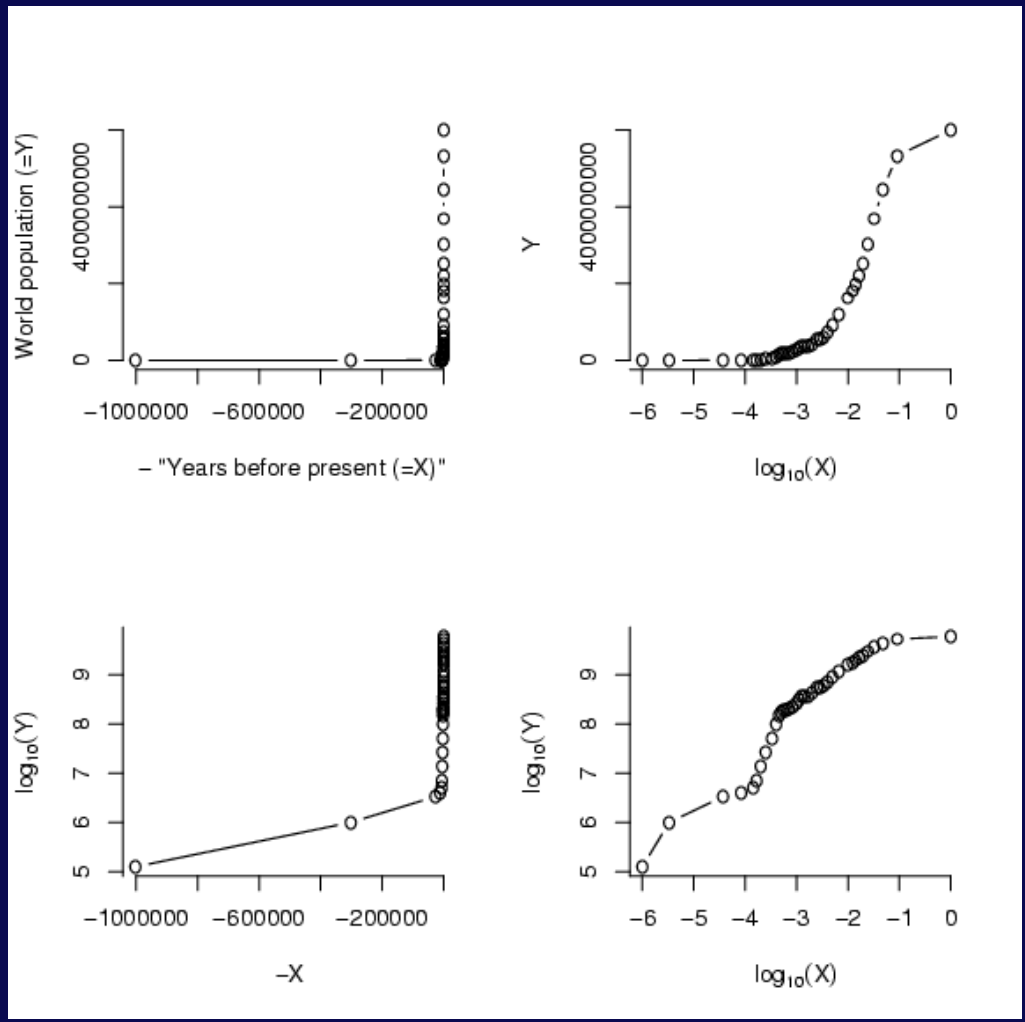
第 12 回：時間の入ったデータの分析

- 時系列データ
  - 時系列でないデータは個々のオブザーベーションが独立。例えば，身長と体重の関係では，Aさんの身長とBさんの身長には関係がないし，Aさんの体重とBさんの体重には関係がない。
  - 時系列で取られているデータは，互いに独立でない。例えば，ある人が生まれてから，毎年誕生日に身長を測って 18 歳くらいまで記録して年齢と身長の関係をみると，8歳のときの身長は，7歳のときの身長がどこまで伸びていたかということにある程度依存する。横軸に西暦年をとり，縦軸に世界人口をとってプロットした場合も同様で，1950 年の人口は，1949 年にどこまで人口が増えていたかということと無関係ではない。
  - 時刻を独立変数にした線形回帰は，よく行われるが，一般には正しくない。微分方程式モデルや差分方程式モデルを立てるか，周期性に着目してスペクトル解析を行う
- 間隔データ：データ数が少ないときに，イベント発生までの間隔を使うと，イベント数だけを使うよりも情報量が多い。ただし，観察の打ち切りも考慮する必要がある。生存時間解析。

時系列データ(1):世界人口の変化

- 普通に散布図を描いてみると指数関数的に見えるが，片対数や両対数でプロットしてみると，そうではなくて3段階くらいの異なるカーブがつながったものであることがわかる。
- 両対数で線形回帰してその線を延ばして元に戻して予測値を出すのは3重の間違い。
- 微分方程式モデルや差分方程式モデルは原理的には悪くないが，世界人口の変化の場合は，農耕革命や産業革命で微分方程式や差分方程式が変化していると考えられるので限界がある。
- シナリオを仮定してシミュレーションをするのが筋。

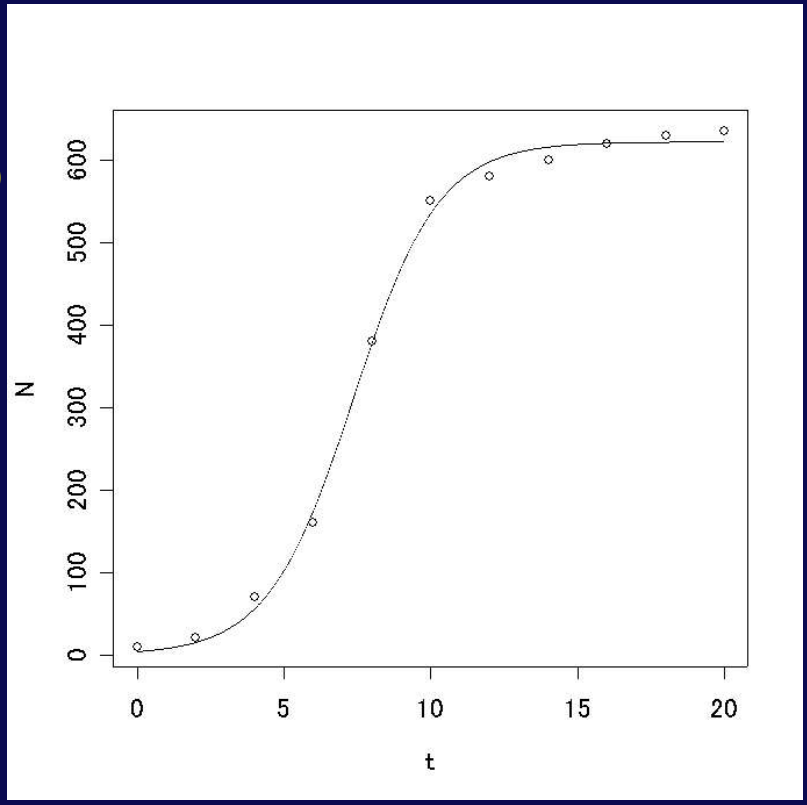


世界人口の変化についての微分方程式または差分方程式モデル

- 微分方程式や差分方程式で隣り合う点の間の関係を説明するアプローチは，時系列データが互いに独立ではないことは正しく反映している
- 未知の係数を求めるには，非線型最小二乗法を用いる。R では nls() を使うか，optim() で関数を最小化するパラメータを求める
- 世界人口の変化についての微分方程式モデル（時刻を t，人口を N，増加率を r，初期人口を N0，人口収容力を K として）
  - 指数的増加モデル：dN/dt=rN，即ち N=N0 exp(rt)
  - ロジスティック増加モデル：dN/dt=rN(1−N/K)
    即ち N=K/{1+(K/N0−1) exp(−rt)}
    ＊ これは，テキストの式と表記法は違うが同値
  - 最後の審判日モデル：dN/dt=rN^2
  - 指数的増加の和のモデル： 2つの部分集団（先進国と途上国）に分けて，それぞれが指数的増加すると考えたもの。先進国の割合を p として，dN/dt=dN1/dt+dp(N−N1)/dt=r1N1+pr2(N−N1)
- 人口は整数なので，微分よりも差分と見るほうが本質的かもしれない。差分方程式ではカオスが起こることもある。

非線形最小二乗法

- 世界人口データでは収束しないので，簡単なサンプルデータでやり方を示す。
- 試験管で培養している酵母菌の量の経時的変化のデータが
  時間　0　2　4　6　8　10　12　14　16　18　20
  量　10　20　70　160　380　550　580　600　620　630　635
  であるとき，酵母菌自身が出す有毒物質が環境抵抗となってロジスティック成長していると考えられるので，それを当てはめてみる。
- P <- data.frame(t=seq(0,20,by=2),N=c(10,20,70,160,380,550,580,600,620,630,635))
  getInitial(N~SSlogis(t,K,tmid,r),data=P)
  res <- nls(N~SSlogis(t,K,tmid,r),data=P)
  summary(res)
  tt <- seq(0,20,by=0.01)
  plot(P)
  lines(tt,predict(res,list(t=tt)))



時系列データの解析（2）

- アプリオリに，そのデータの変化のパタンがいくつかの成分によって構成されると決めてしまい，その中身を探るアプローチ。
- 時系列データには，繰り返し起こる変化を含むように見えるものがある。例えば，日本のような中緯度に位置する場所では，一日の平均気温は，季節ごとに周期的に変化する。しかし，繰り返しは完全ではなく，地球温暖化が起これば長期的には上昇傾向をもつし，天気などによって毎日微妙に変化する。このような時系列データは，季節成分，傾向成分，不規則成分という3つの成分に分解して考えることができる。
- 自己回帰（AR）モデルの考え方：適当なタイムラグ s をおいて，周期的に同じ成分によって決まる値が出現するなら，x(t) と x(t+s) は相関するはず。x(t+s) ~ x(t) というモデルが充分にデータに当てはまればいい。当てはまりは AIC などで評価
- 正弦波と余弦波の和に分解するスペクトル解析では，フーリエ変換やウェーブレット変換が良く使われる。

気温データへの AR の当てはめ

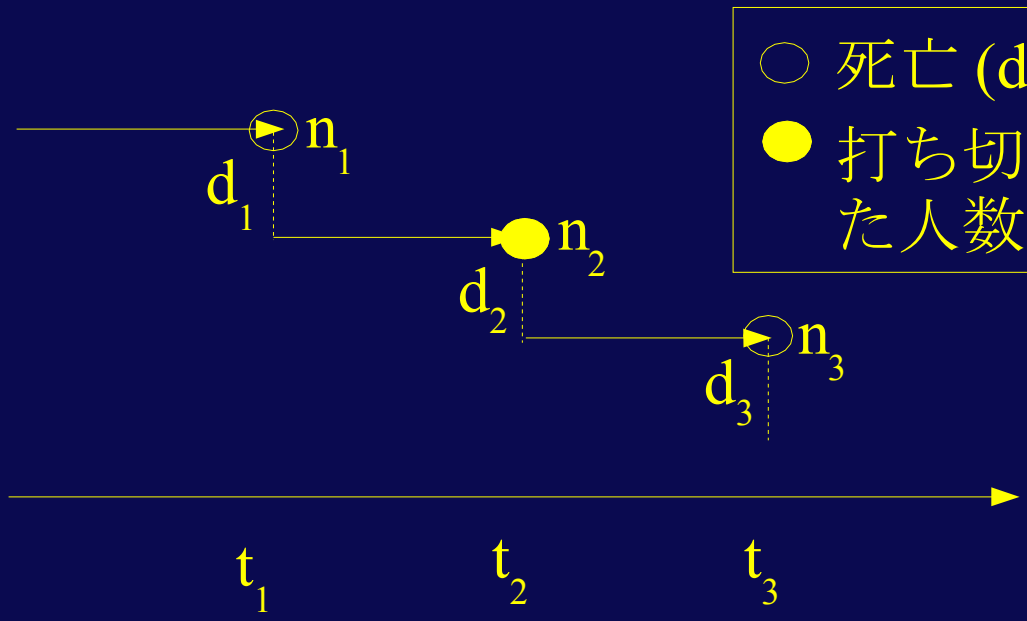
- ソロモン諸島ガダルカナル島で 1995 年 11 月 23 日から 12 月 29 日まで毎日朝 6 時と昼 2 時の 2 回ずつ気温を測定したデータがある。
- これを順番に temp という変数に代入するならば,
temp <- c(23.5, 28.7, 24.4, 28.5, 24.5, 30.5, 25.0, 30.9, 25.0, 26.7,
24.1,30.3, 25.4, 28.3, 24.5, 32.9, 26.0, 29.4, 25.7, 31.2, 24.9, 29.3, 24.6,
29.9,24.8, 32.0, 26.3, 31.8, 24.2, 31.2, 24.7, 29.6, 24.8, 30.7, 25.8, 31.7,
25.4,30.1, 24.8, 29.1, 25.4, 30.9, 23.6, 31.2, 26.1, 30.8, 24.9, 32.2, 26.2,
31.2,25.1, 31.9, 25.8, 32.5, 24.8, 32.4, 25.3, 31.7, 25.8, 33.6, 25.5,
31.6,26.0, 30.5, 25.0, 33.0, 25.5, 30.0, 23.5, 31.6, 25.9, 33.0, 24.8, 33.3)
- 時系列データと認識させるため,
library(ts) としてから
ttemp <- ts(temp, freq=2, start=c(23, 1)) として, 毎日 2 回の測定値があるデータを 23 日の 1 回目から, temp というベクトルから読んで, ttemp という時系列データに付値するという操作を意味する。
- ttemp に対して自己回帰モデルを当てはめるには, ar(ttemp) とすればいい。

生存時間データの解析

- 間隔データ（イベントが起こるまでの期間のデータ）には打ち切りが生じやすい
- まだイベントが起こっていない割合の経時的な変化を生存関数（生存曲線）と呼ぶ
- 大標本なら生命表解析を行う
- パラメトリックな分布関数を当てはめる方法もある（加速モデルと呼ばれる）
- R では，survival ライブラリに含まれるので，library(survival) としないと使えない
- 他の変数が生存時間に与える影響を分析：比例ハザードモデル：coxph() 関数
- 生存曲線から半数生存時間（メディアン）を求める：基本はカプラン・マイヤ法：survfit() 関数
- 生存曲線の差を検定：ログランク検定など：survdiff() 関数

カプラン・マイヤ法

- イベントが起こる可能性のある人口を分母，まだイベントが起こっていない人口を分子として，時間の長さを横軸にして，イベントが起こった場合は分子から1を引き，打ち切りが起こった場合は分母からも分子からも1を引くという形で，時間と人数の積和（Kaplan–Meier の積・極限推定量）がちょうど全体の半分になるまでの時間をイベントが起こるまでの時間の中央値とする考え方。
- 死亡データの場合は下図のように捉えられる。



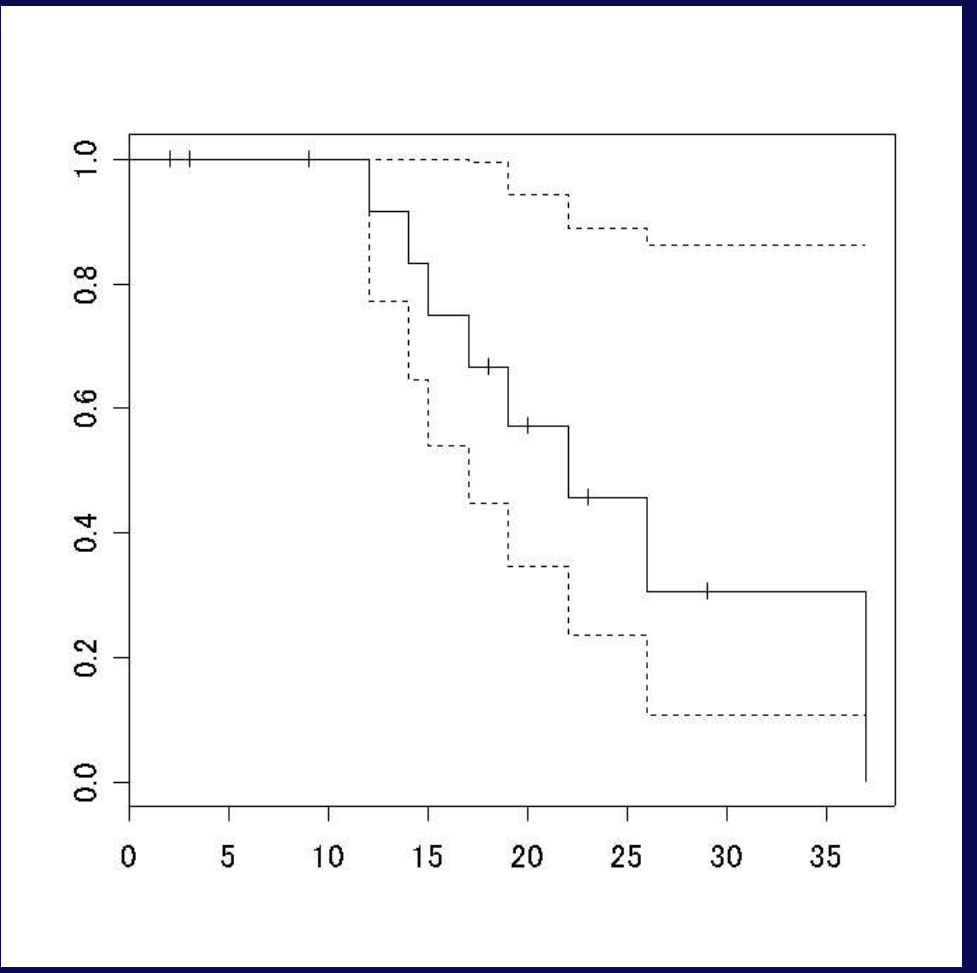
○ 死亡 (d は死亡した人数)

● 打ち切り (d は打ち切りが起こった人数)

$(1-d_1/n_1)(1-d_3/n_3)\ldots$ と掛けていき，初めて 0.5 を下回る時点 t が中央値の点推定量となる

カプラン・マイヤ法の計算

- ソロモン諸島のある村で，すべての母親に対して出産暦聞き取りを行って得られたデータで，第1子出生が 1986 年以降のものの第1出産間隔データを R で分析するには，以下のプログラムでよい。




```
time<-c(17,14,22,37,12,15,19,26,29,23,20,18,9,9,3,2)
#  第2子出産に達したレコードを１，打ち切りを０とする
event<-c(1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0)
library(survival)
dat<-Surv(time,event)
res<-survfit(dat)
summary(res)
plot(res)
```