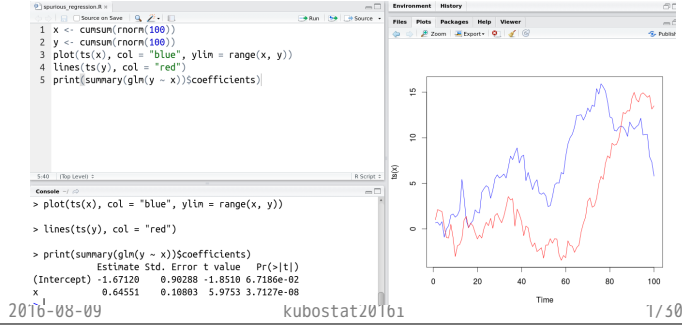


時系列データ解析 状態空間モデル (SSM) の続きと 疑わしい回帰 (spurious regression)

久保拓弥 (北海道大・環境科学)



今回、説明してみたいこと

- 時系列データ: 単純な回帰はダメ(続)
- 状態空間モデル: 乱歩と雑音の分離
- 欠測と不等間隔
- 時系列「ばらばら解析」やめよう
- 「うたがわしい回帰」への対策

階層ベイズモデル!

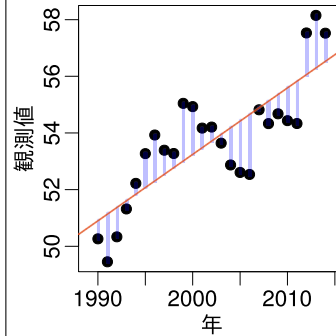
2016-08-09

2/30

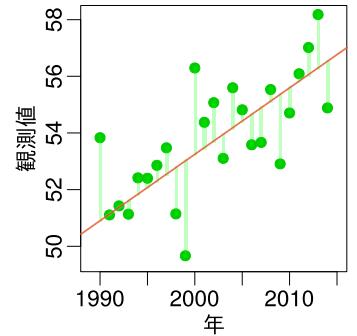
今日の要点

時系列データの解析は
階層ベイズモデル化した
状態空間モデルを使うのが便利

時系列の「ずれ」



GLM のずれ



直線からのずれがちがう!

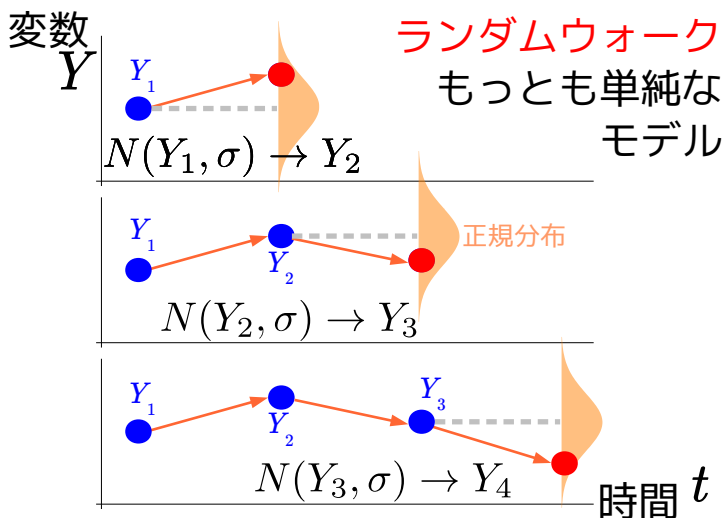
時間的自己相関がある

時間的自己相関がない

2016-08-09

kubostat2016i

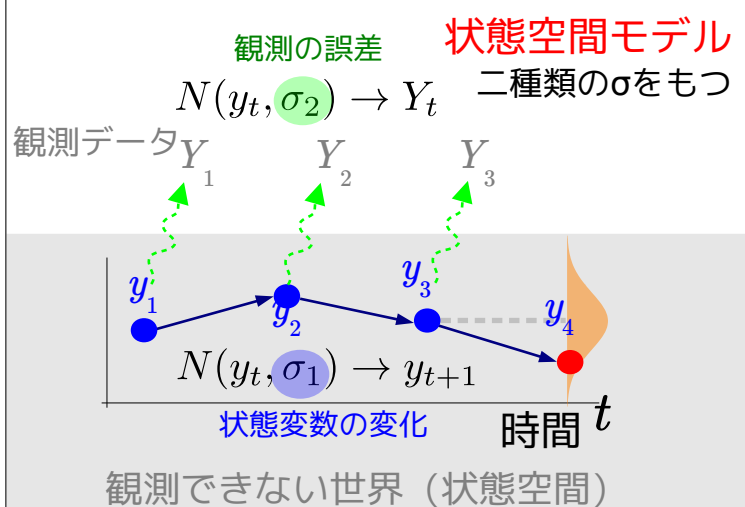
4/30



2016-08-09

kubostat2016i

5/30



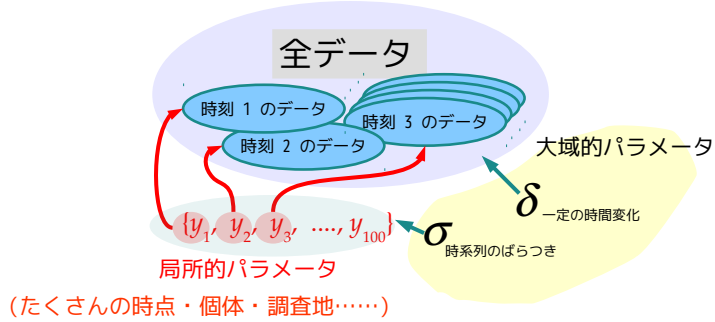
2016-08-09

kubostat2016i

6/30

状態空間モデルは階層ベイズモデル

多数の「似たようなパラメーター」たちに「適切」な制約を加えて推定できる



2016-08-09

kubostat2016i

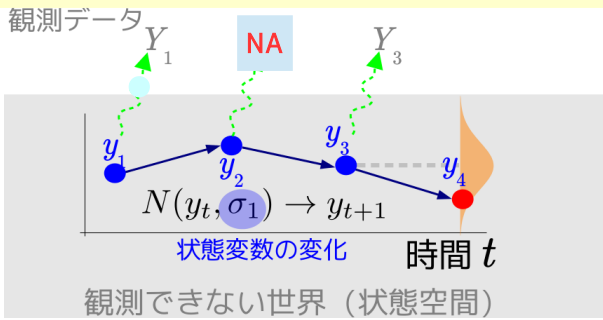
7/30

状態空間モデルを使う利点

欠測とか不等間隔とか

状態空間モデル + 観測モデル

欠測があっても問題ない
「補完」の必要なし!

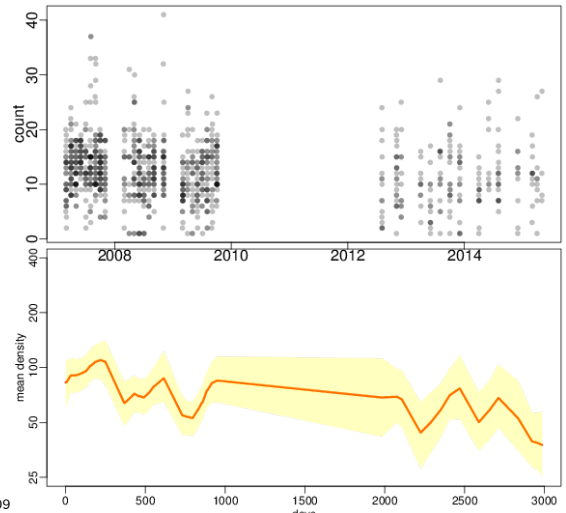


2016-08-09

kubostat2016i

9/30

不等間隔データでも何とかできます!



2016-08-09

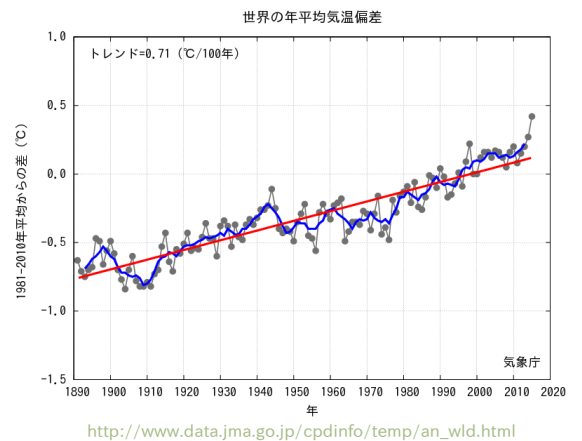
10/30

状態空間モデルを使う利点

「ばらばら解析」の回避

気象庁のデータ解析?

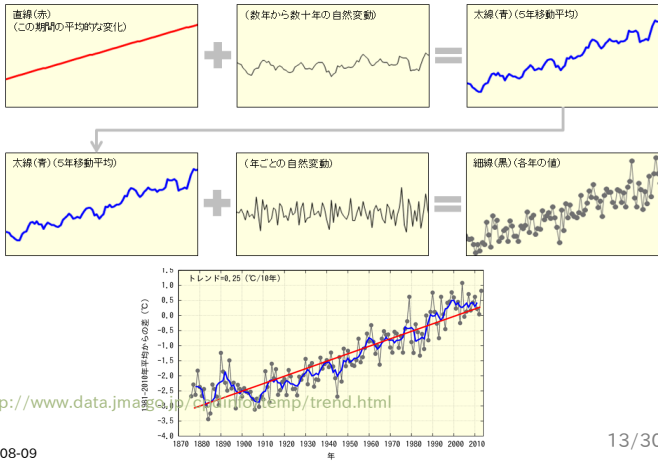
気象庁の長期変化傾向 (トレンド) の解説



2016-08-09

12/30

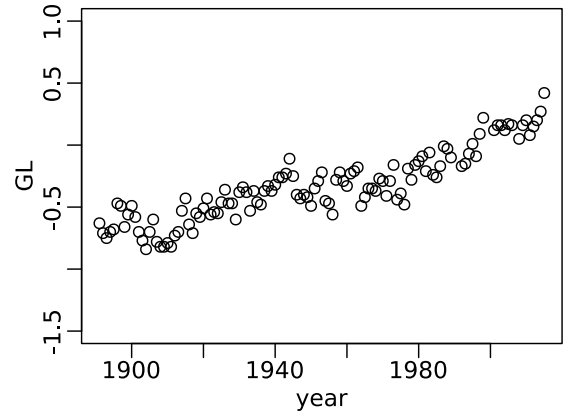
気象庁の長期変化傾向（トレンド）の解説



2016-08-09

13/30

公開データをダウンロード



2016-08-09

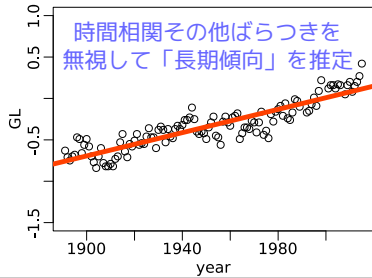
14/30

「とりあえず、直線回帰」の危険性

```
> summary(glm(GL ~ year, data = d))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.41e+01	6.21e-01	-22.6	<2e-16
year	.7.03e-03	3.18e-04	22.1	<2e-16



2016-08-09

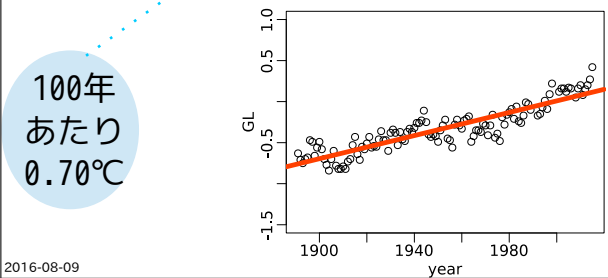
15/30

直線あてはめ (GLM) が予測した「温暖化」

```
> summary(glm(GL ~ year, data = d))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.41e+01	6.21e-01	-22.6	<2e-16
year	.7.03e-03	3.18e-04	22.1	<2e-16

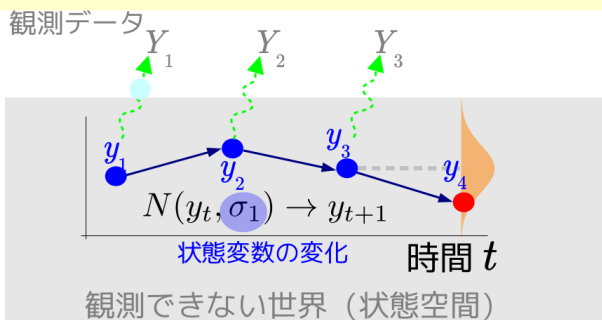


2016-08-09

16/30

状態空間モデル：すべてを同時に推定

ランダムウォーク+各年独立なノイズ



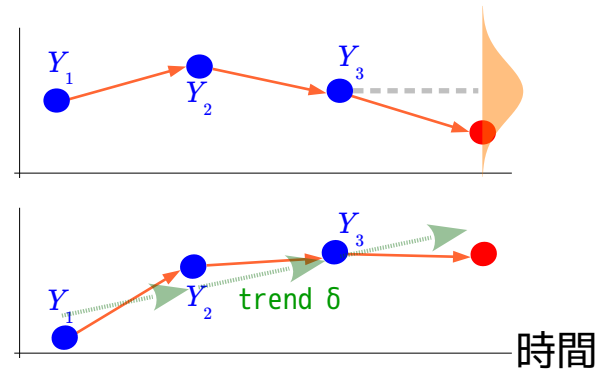
2016-08-09

kubostat2016i

17/30

状態空間モデル：すべてを同時に推定

ランダムウォーク+各年独立なノイズ



2016-08-09

kubostat2016i

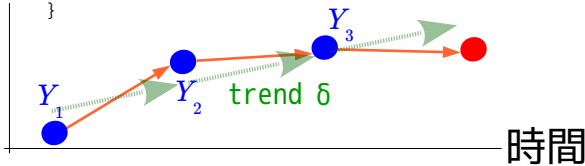
18/30

状態空間モデル：すべてを同時に推定

```

Y[1] ~ dnorm(y[1], tau[2])
y[1] ~ dnorm(0.0, Tau.Noninformative)
for (t in 2:N.Y) {
  Y[t] ~ dnorm(y[t], tau[2])
  y[t] ~ dnorm(m[t], tau[1])
  m[t] <- delta + y[t - 1]
}
delta ~ dnorm(0, Tau.Noninformative)
for (k in 1:2) {
  tau[k] <- 1.0 / (s[k] * s[k])
  s[k] ~ dunif(0, 1.0E+4)
}

```

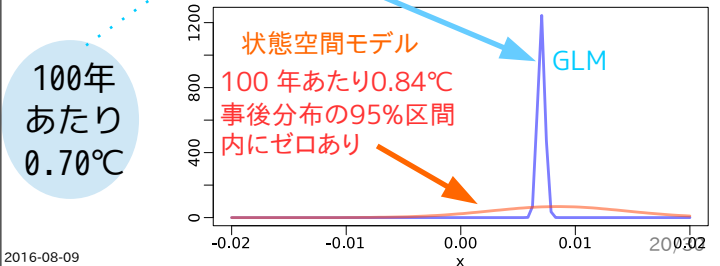


状態空間モデルが予測した「温暖化」

```
> summary(glm(GL ~ year, data = d))
```

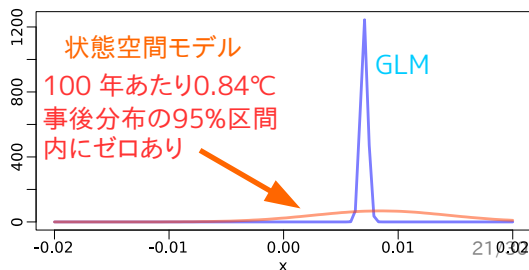
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.41e+01	6.21e-01	-22.6	<2e-16
year	7.03e-03	3.18e-04	22.1	<2e-16



観測値間に相関あり→サンプルサイズが小さくなる

100年
あたり
0.70°C



疑わしい回帰
spurious regression

時系列どうしの回帰
time series $Y \sim$ time series X

時系列データの統計モデリング

でやめたほうがいいこと

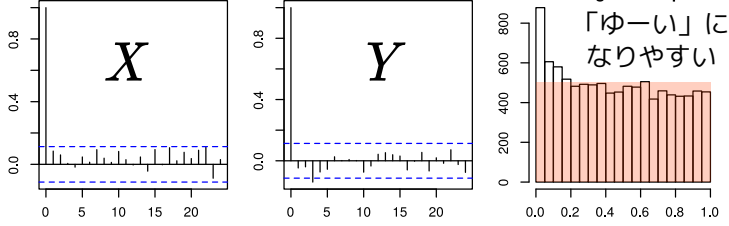
- ・ GLM: $Y(t) \sim t$ とか $Y(t) \sim X(t)$
- ・ 段階的解析: 観測値の四則演算
- ・ 「残差」の再解析
- ・ 「対応」の無視 - 再測は時系列

「見せかけの回帰」 spurious regression

ちょっとだけ実演してみます

ノイズの大きな時系列にうもれたワナ？

時間的自己相関のない時系列？



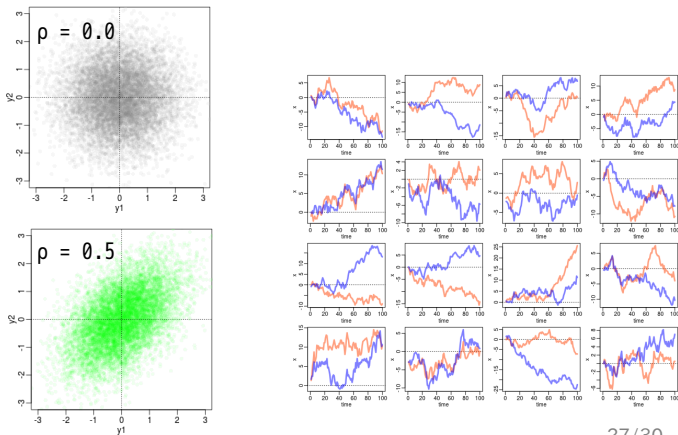
Histogram of v.p1
「ゆーい」に
なりやすい

しかし glm(Y ~ X) とすると...

疑わしい回帰
spurious regression

状態空間モデル (SSM) で
あつかえないか？

二変量正規分布とランダムウォーク



二変量正規分布を部品とする状態空間モデル

```
for (i in 1:N.Y) {
  Y[i, 1:2] ~ dnorm(mu[1:2], Omega[1:2, 1:2])
}
mu[1] ~ dunif(-1.0E+4, 1.0E+4)
mu[2] ~ dunif(-1.0E+4, 1.0E+4)
Omega[1:2, 1:2] <- inverse(VarCov[1:2, 1:2])
VarCov[1, 1] <- sigma[1] * sigma[1]
VarCov[1, 2] <- sigma[1] * sigma[2] * rho
VarCov[2, 1] <- sigma[2] * sigma[1] * rho
VarCov[2, 2] <- sigma[2] * sigma[2]
sigma[1] ~ dunif(0.0, 1.0E+4)
sigma[2] ~ dunif(0.0, 1.0E+4)
rho ~ dunif(-1.0, 1.0)
```

(R で実演)

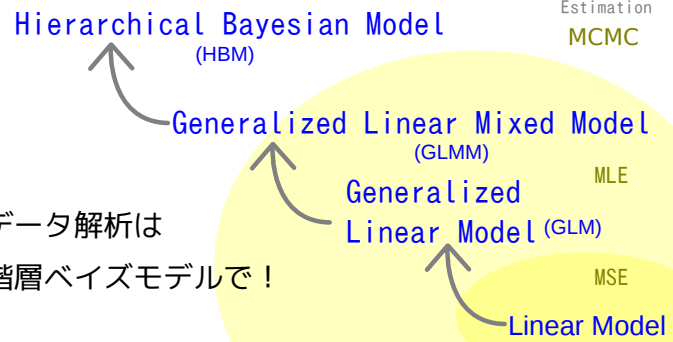
階層ベイズモデルである
状態空間モデル
から得られた事後分布

```
3 chains, each with 5200 iterations (first 200 discarded)
n.sims = 15000 iterations saved
      mean   sd  2.5%  25%  50%  75%  97.5%  Rhat  n.eff
mu[1] -0.122 0.110 -0.342 -0.195 -0.120 -0.048 0.090 1.001 6000
mu[2] -0.157 0.100 -0.355 -0.224 -0.157 -0.091 0.041 1.002 1500
sigma[1] 1.091 0.079 0.949 1.036 1.086 1.142 1.261 1.001 6100
sigma[2] 0.993 0.074 0.864 0.941 0.987 1.039 1.151 1.001 4100
rho      0.568 0.070 0.420 0.523 0.573 0.617 0.693 1.001 11000
```

ふたつの時系列データの変動が
関連しているかどうかを特定できる

おしまい

The Evolution of Linear Models



データ解析は
階層ベイズモデルで！