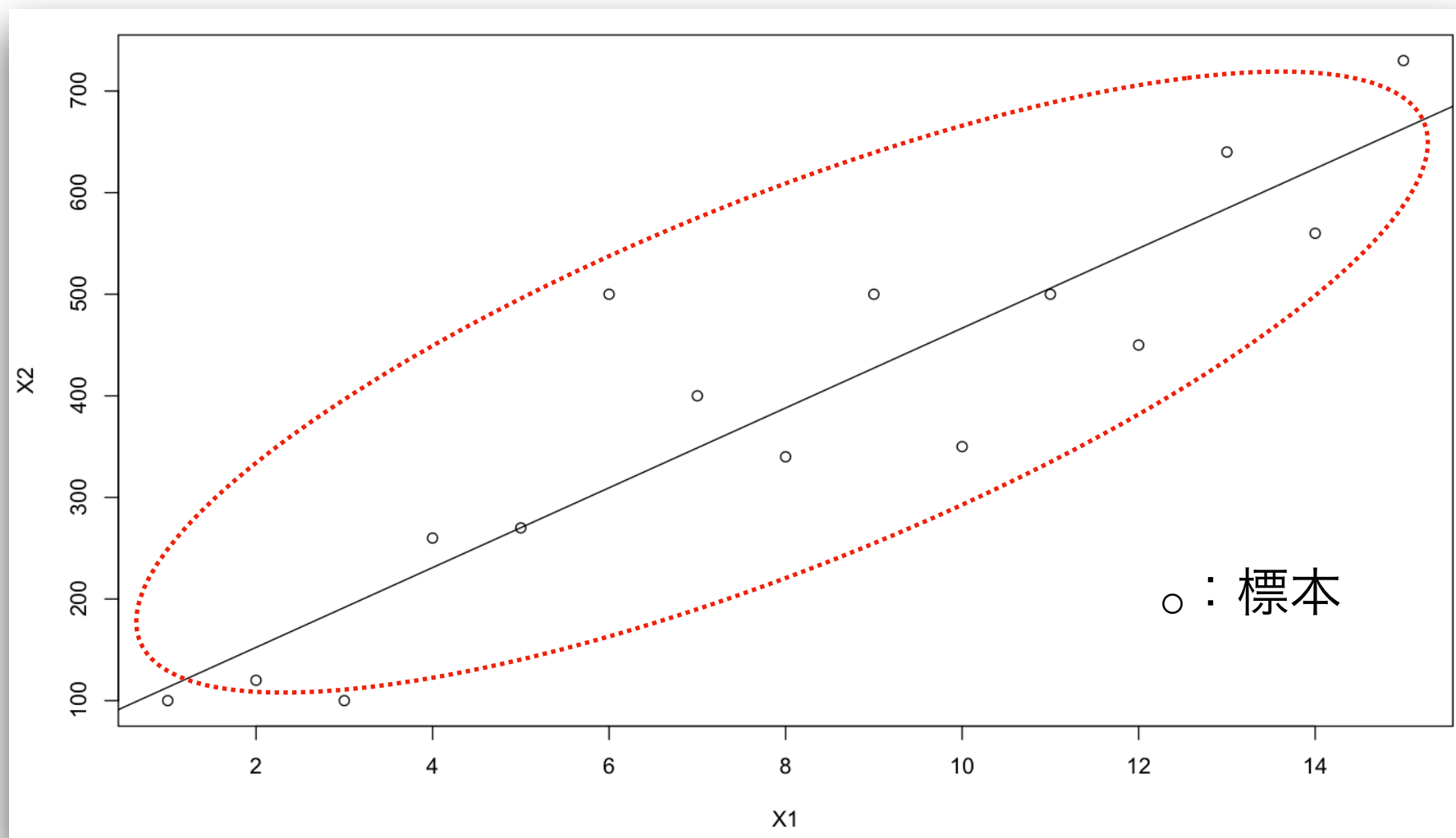


多変量自己回帰モデルとの比較を通じた 遅延相関分析の有用性評価

蘆田 誠人

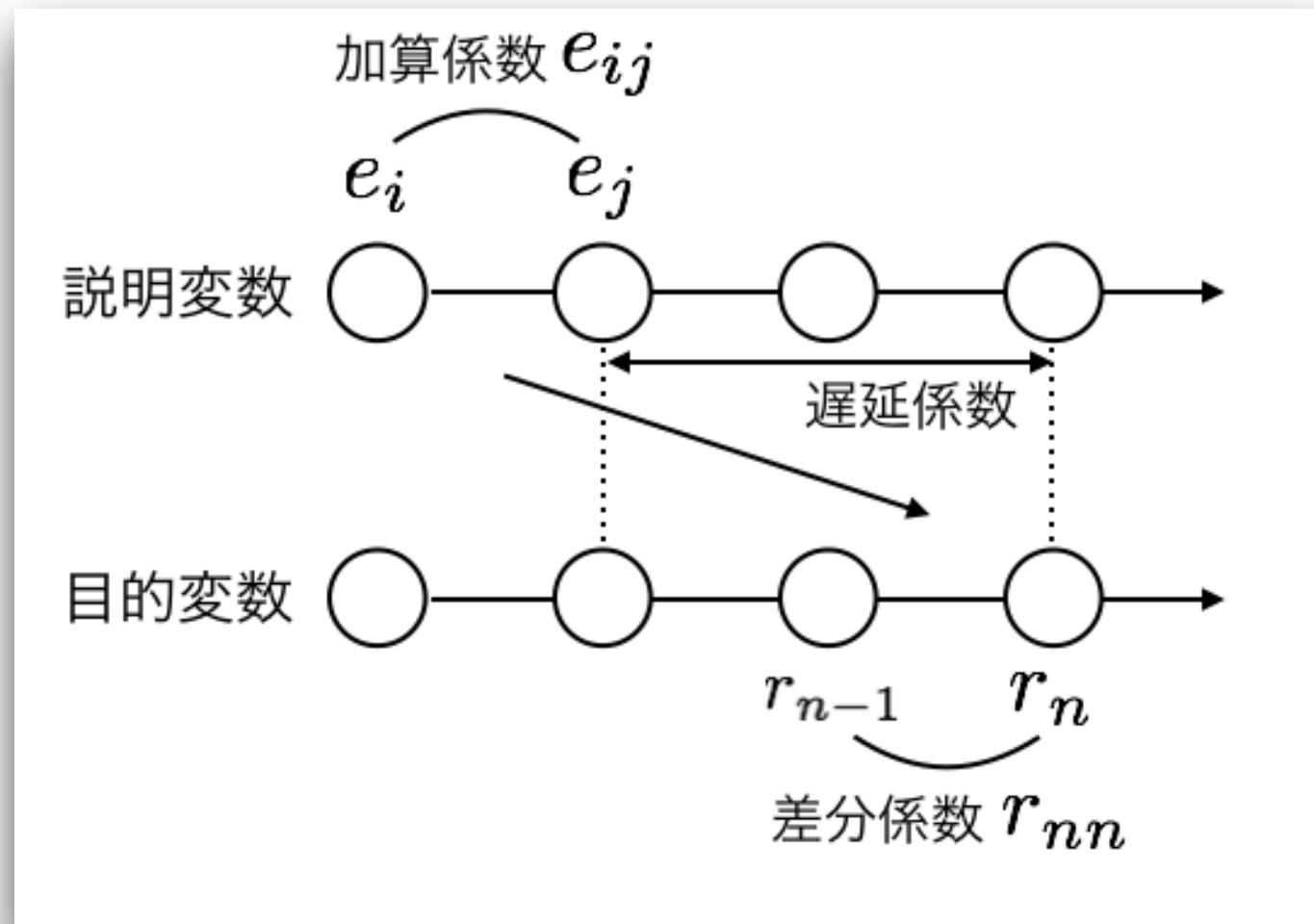
背景

- 相関分析とは
 - ▶ 二変数間に直線的な関係があるかどうか



背景

- 遅延相関分析
 - ▶ 時間的遅延を考慮した相関分析手法
 - ▶ 三つの係数を基に相関分析を行う



先行研究

- 生活習慣から健康状態への影響の調査
 - ▶ 生活習慣がどの程度の遅延を経て健康状態に影響を及ぼすか
 - ▶ 生活習慣と健康状態の関係が時間を遅延させたことにより抽出できた
- OSSの共進化の調査
 - ▶ システムとコミュニティの共進化を定量化
 - ▶ 従来の相関分析では見れなかった時間遅延に基づく関係が抽出できた

目的とアプローチ

- 目的

- ▶ 遅延相関分析の有用性を評価すること

- アプローチ

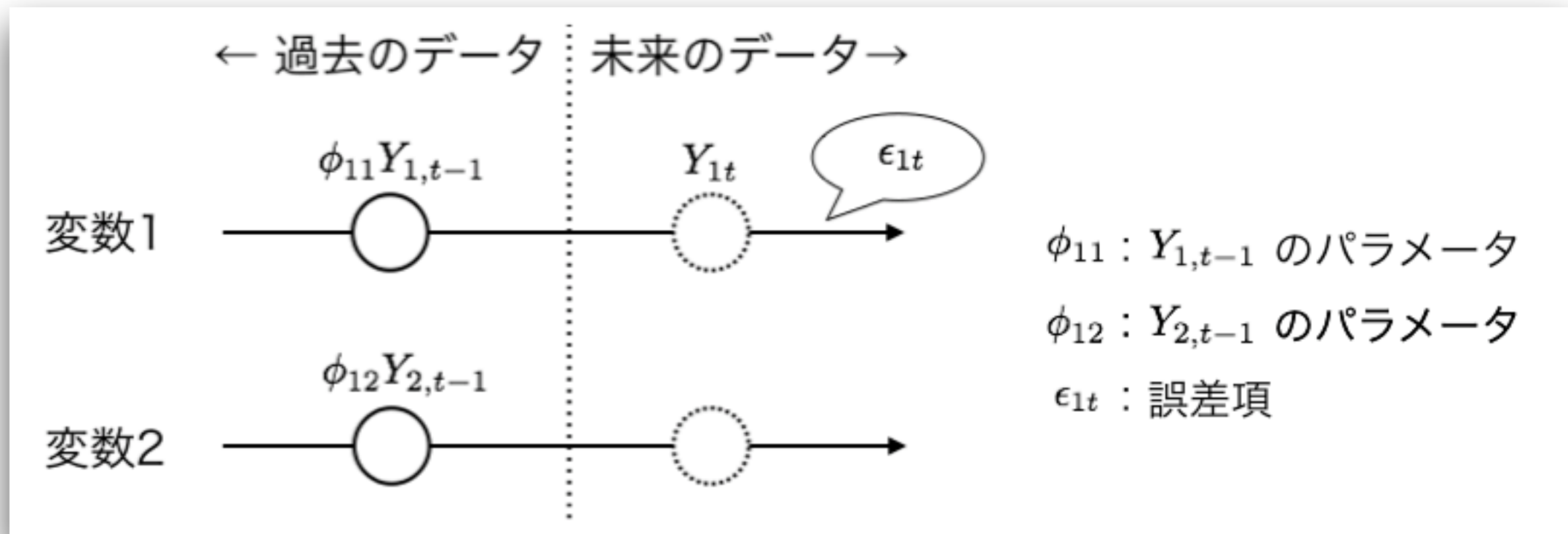
- ▶ 時系列モデルと比較する

- 1.遅延相関分析の遅延係数を用いて，回帰モデルを構築
- 2.回帰モデルと時系列モデルのそれぞれの予測誤差で比較

時系列モデル

- 過去のデータから現在のデータを予測するモデル
 - 多変量自己回帰(VAR)モデル

$$Y_{1t} = \Phi_{11}Y_{1,t-1} + \Phi_{12}Y_{2,t-1} + \epsilon_{1t}$$



定常状態

- 分散が無限大のデータを用いてモデル適用する
 - ▶ 予測値に大きな誤差が生まれる

定常データの例

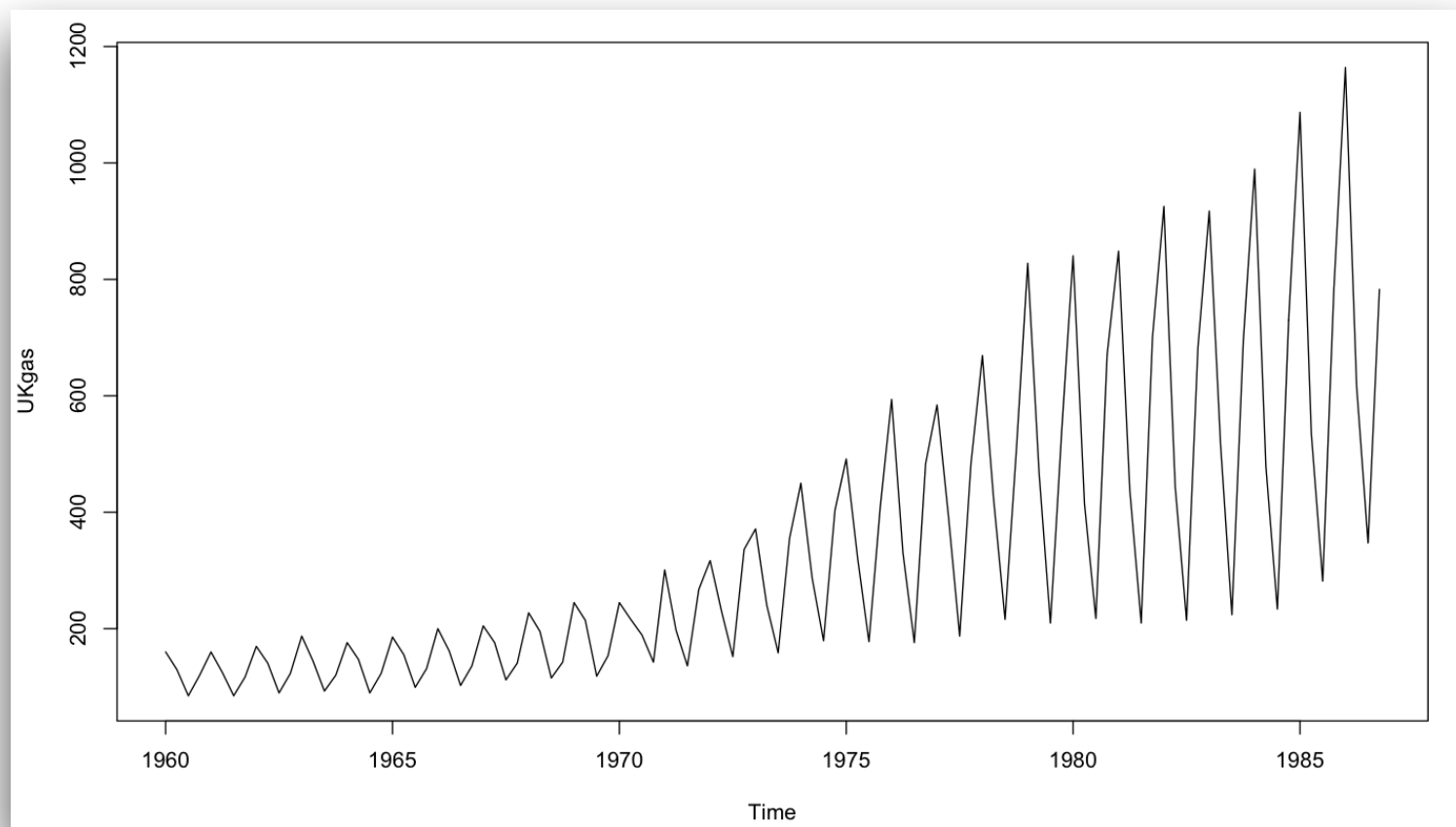
✓ サイコロの出目：周囲の条件が変更されても確率は不変

非定常データの例

✓ 日本の外気温：季節によって朝夕の気温差は変動

定常化

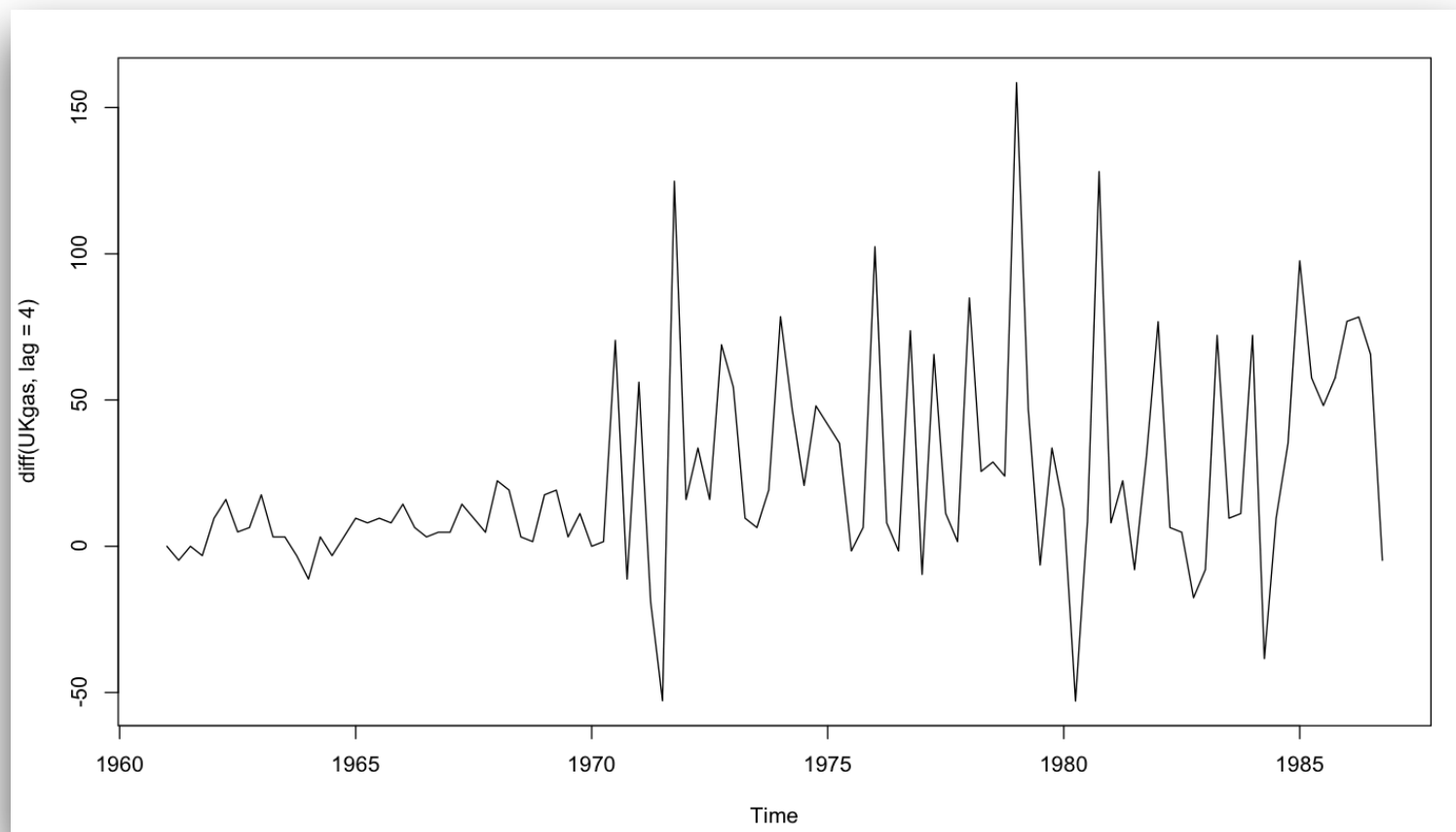
- ・ イギリスのガス消費量(1960年~1986年)
 - ▶ 生データ (縦軸：ガス消費量(サーム), 横軸：年)



- ・ 上昇トレンド
 - ・ 分散が大きくなる
- 定常条件を満たしていない**

定常化

- 上昇トレンドの除去・季節性の除去
 - ▶ 差分を取り，ラグを取ることでより定常な状態



- トレンド無し
 - 分散はほぼ一定
- 定常条件を満たした

遅延相関分析

- 三つの係数により構成される

- ▶ 加算係数(1~12)

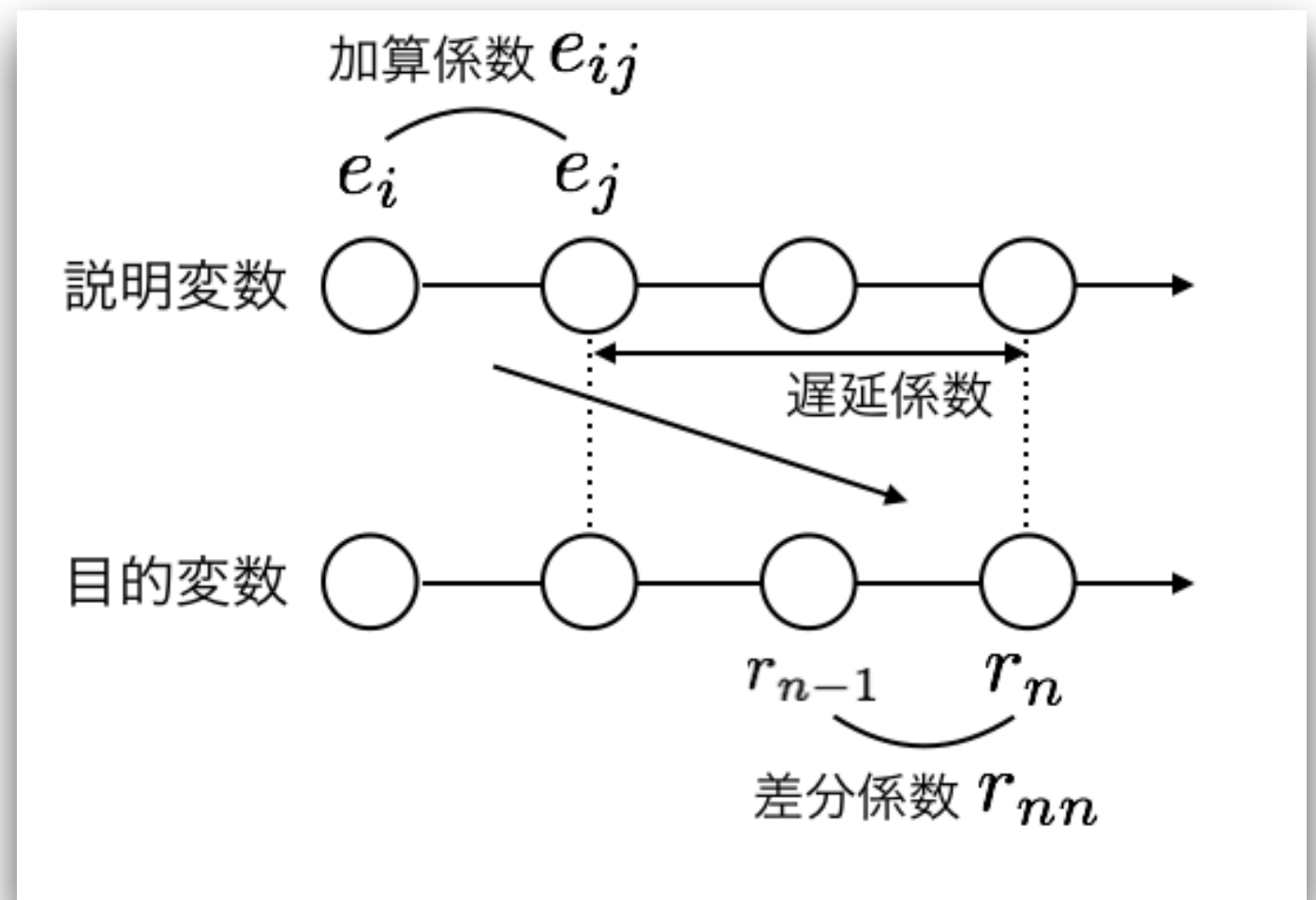
- 説明変数の累積値

- ▶ 差分係数(1~12)

- 目的変数の変化量

- ▶ 遅延係数(1~12)

- 遅延した期間



遅延相関分析

- 最大相関係数

▶ 三つの係数のセットで、相関が一番高いセットを記憶

第一メトリクス	第二メトリクス	相関係数	加算	差分	遅延
commitNum	countLine	0.45	2	3	1
countLine	totalFixers	0.23	10	2	5
maxCyclomatic	totalFixed	-0.13	8	2	2
totalReplies	totalReporters	0.39	4	1	10

実験方法

- 遅延係数を用いた回帰モデル
 - ▶ 目的変数を遅延係数分ずらしたデータで回帰モデルを構築

	説明変数	目的変数
t_1	2	7
t_2	3	8
t_3	4	9
t_4	5	10
t_5	6	11

遅延係数分ずらす

	説明変数	目的変数
t_1	2	8
t_2	3	9
t_3	4	10
t_4	5	11
t_5	6	12

データセット

- Apache HTTP Serverプロジェクト
 - ▶ 2001年から2015年のデータからメトリクスを抽出

メトリクス	件数
ソースコード	39
不具合修正履歴	41
メーリングリスト	10

比較方法

- データセットを分割
 - ▶ 学習用データセット：2001年1月～2010年12月
 - ▶ テスト用データセット：2011年1月～2015年12月
- テスト用データと各モデルの予測値を相対誤差で比較

$$RE = \frac{|PredictedValue - MeasuredValue|}{MeasuredValue}$$

RE ：相対誤差 $PredictedValue$ ：予測値 $MeasuredValue$ ：実測値

実験結果

- 相関係数の絶対値が0.4以上の関係について各モデルで予測
 - ▶ 349件/8010件が抽出された

	精度が高かった数	誤差平均	誤差の中央値	最大誤差
VARモデル	191	0.3997	0.3212	1.8714
回帰モデル	26	0.7997	0.6155	4.8651
N/A	132	-	-	-

考察

- 回帰モデルの方が精度が良かった26件について

まとめ