ソフトウェア工学III
プロジェクト特性データの分析III
——予測・ルール発見——

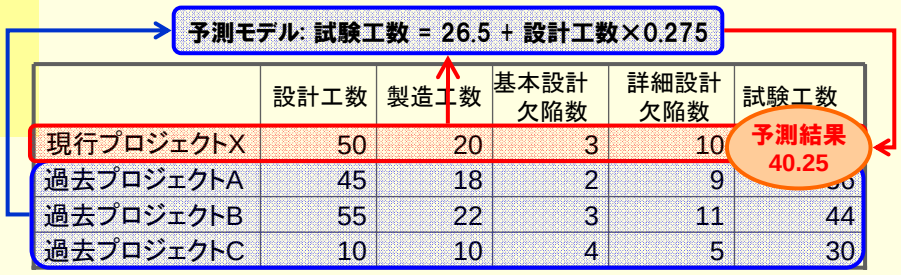
ソフトウェア工学講座
門田暁人
akito-m@is.naist.jp
B303室，内線5311

分析の目的

- データ間の関係を調べる.
  - 視覚的に
    - 散布図，ヒストグラム，箱ひげ図，平行座標プロットなど
  - 定量的に
    - t検定，カイ二乗検定，分散分析，無相関検定など
    - 相関係数，クラメールのV，回帰曲線など
- データの予測（見積もり）を行う.
  - 重回帰分析，協調フィルタリング，マハラノビスタグチ法など
- 大量のデータの中から隠された関係を発見する.
  - アソシエーション分析（相関ルール分析）

プロジェクト特性データに基づく見積もり

- 過去プロジェクトのデータから予測モデルを作成する.
  - 線形モデル(重回帰分析), ニューラルネットなど
- 現行プロジェクトの実績値を予測モデルに当てはめ, 工数, バグ数, などを予測する.



	設計工数	製造工数	基本設計欠陥数	詳細設計欠陥数	試験工数
現行プロジェクトX	50	20	3	10	
過去プロジェクトA	45	18	2	9	
過去プロジェクトB	55	22	3	11	44
過去プロジェクトC	10	10	4	5	30

何を見積もる(予測する)のか？

- 開発工数(人月もしくは人時)
  - 開発総工数
  - テスト工数
- 出荷後品質(バグ数, バグ密度)
  - モジュール単位
- プロジェクトの成否(コスト超過, 納期超過,)
  - 失敗 or 成功

工数, コスト,

コスト見積もり手法（ソフトウェア工学II 参照）

- 契約価格に基づく決定
- パーキンソン（Parkinson）の法則
- 専門家による判定
- 積算法
- 計算式等のコストモデルによる算出
- 類似プロジェクトからの類推

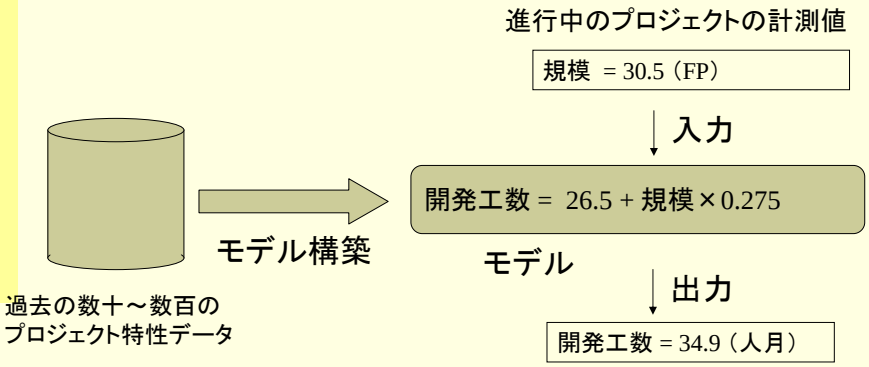
プロジェクト特性データを利用可能

プロジェクト特性データに基づく見積もり

- 定義済みモデルに基づく見積もり
  - COCOMO，COCOMO II，Agile COCOMO
- 過去のデータに基づく見積もり
  - モデルベース手法
    - 重回帰分析，CoBRA法，ニューラルネット，．．．
  - メモリベース手法
    - Analogy-based法，協調フィルタリング法，OSR法

モデルベース予測

- 重回帰分析，CoBRA法，ニューラルネット，．．．



重回帰分析（線形回帰モデル，重回帰モデル）

- モデル式

$$\hat{Y} = a_1 N_1 + a_2 N_2 + \cdots + a_k N_k + C$$

$\hat{Y}$：従属変数（目的変数）の予測値
N_j：独立変数（説明変数）
a_j：係数（偏回帰係数）
C：定数項

実測値 Y と予測値 $\hat{Y}$ の差を残差と呼ぶ．
残差の2乗和が最小となるように a_j と C を定める．

仮定１：各説明変数は，互いに独立である．
仮定２：目的変数は，正規分布に従う．
仮定３：各説明変数と目的変数は直線相関関係にある．

重回帰モデルは工数予測モデルとして妥当か？

- モデル式

$$\hat{Y} = a_1 N_1 + a_2 N_2 + \cdots + a_k N_k + C$$

仮定1：各説明変数は，互いに独立である．
→規模（FP），工期，開発要員数など，独立とはいえない．
→多重共線性がある．
重回帰モデルの説明変数間に強い関連が存在することにより，モデル式が構築できなかったり，予測結果に信頼が置けなくなる現象．

解決策
→変数選択を行う．（もしくは，変数の合成を行う）

変数選択

- 目的
  - 多重共線性を避ける．（互いに相関の強い説明変数が重回帰モデルに取り入れられないようにする）
  - 予測に効いてない説明変数を除去する．
- 基本方針
  - 予測に有用な説明変数のみを重回帰モデルに取り入れる．
    a. 偏回帰係数の有意性検定に基づく（ゼロでない＝有意）
    b. 「モデルの良さ」の基準を用いる．
       - 赤池の情報量基準（AIC: Akaike's Information Criterion）
       - 自由度調整ずみ重相関係数 など
    c. 実際に予測してみて判断する．（判断のためのデータセットが必要）
- 留意点
  - （どんな変数であれ）説明変数の数を増やすと残差は小さくなる．
    →残差（残差平方和）は，変数選択の基準になり得ない．

変数選択法

- 総当たり法
  - 昔は現実的ではなかった．コンピュータの発達した現在は有力である．
- ステップワイズ法
  - 変数増加法
    - 1つずつ説明変数を重回帰モデルに取り入れていく
  - 変数減少法
    - 全説明変数を使った回帰モデルから，1つずつ説明変数を削除していく
  - 変数増減法（狭義のステップワイズ法）
    - 基本的には1つずつ説明変数を追加していくが，過去に追加したものの中に除去すべき説明変数がないかチェックする．
  - 変数減増法
    - 基本的には1つずつ説明変数を削除していくが，過去に削除したものの中に取り入れるべき説明変数がないかチェックする．
- 直行表を用いた方法

重回帰モデルは工数予測モデルとして妥当か？

- モデル式

$$\hat{Y} = a_1 N_1 + a_2 N_2 + \cdots + a_k N_k + C$$

仮定2：目的変数は，正規分布に従う．
→工数は値の小さい部分に偏っている．正規分布とはいえない．

解決策
→対数変換を行う．
工数の代わりに$\log_{10}$(工数)を使う

重回帰モデルは工数予測モデルとして妥当か？

- モデル式

$$\hat{Y} = a_1 N_1 + a_2 N_2 + \cdots + a_k N_k + C$$

仮定3：各説明変数と目的変数は直線相関関係にある．
→直線相関関係にあるとはいえない．
一般に，規模が大きくなると工数は指数的に増大する．

解決策
→指数曲線回帰を使う．

$$\hat{Y} = C N_1^{a_1} N_2^{a_2} \cdots N_k^{a_k}$$

両辺対数取ると

→説明変数，目的変数共に対数変換してから重回帰分析する．

$$\log \hat{Y} = b_1 \log N_1 + b_2 \log N_2 + \cdots + b_k \log N_k + C_0$$

これを「対数線形モデル」と呼ぶ．

モデルに対する言い訳

- 工数が

$$\hat{Y} = a_1 N_1 + a_2 N_2 + \cdots + a_k N_k + C$$

のような式で表現できるという理論的根拠はない．

- 有名な言葉：
  - All models are wrong. Some of them are useful.
  - 予測精度が高ければ役に立つ．
- 人間が介在する以上，厳密なモデル化は難しい．また，厳密にモデル化することが望ましいともいえない．現実的には，「簡潔さ」「説明変数の計測の容易さ」「高い予測精度」が求められる．

例題1

- canada.csvにおいて，
  - 目的変数
    - ActualEffort
  - 説明変数
    - Duration，ExpEquip，ExpProjMan，Adj FPs，Dev Env
- として重回帰モデルを作成せよ．

- JavaScript による重回帰分析
  - 群馬大学社会情報学部の青木繁伸教授による
    - http://aoki2.si.gunma-u.ac.jp/JavaScript/mreg.html

例題1の結果

変数	偏回帰係数	標準誤差	t値	P値	標準化偏回帰係数
Var01	146.1377	63.42289	2.30418	0.02414	0.2367945
Var02	-269.9088	261.6748	1.03147	0.30582	-0.08561447
Var03	290.0082	226.9117	1.27807	0.20539	0.1052341
Var04	13.06223	2.310043	5.65454	0.00000	0.5812246
Var05	-1894.870	441.4975	4.29192	0.00006	-0.3240460
定数項	2299.255	971.3267	2.36713	0.02066	

重回帰モデル式

ActualEffort =

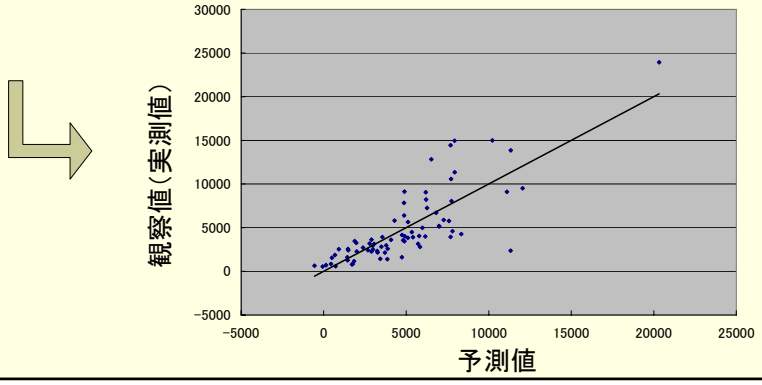
146.1377*Duration – 269.9088*ExpEquip + 290.0082 * ExpProjMan
+ 13.06223 * Adj FPs – 1894.870*Dev Env + 2299.255

例題1の結果

予測結果

番号	観察値	予測値	残差	標準化残差
1	5152.000000	6992.955093	-1840.955093	-0.751814
2	5635.000000	5103.539044	531.460956	0.220733
3	805.000000	1715.085434	-910.085434	-0.375773
4	3829.000000	5092.929905	-1263.929905	-0.521641

. . . .



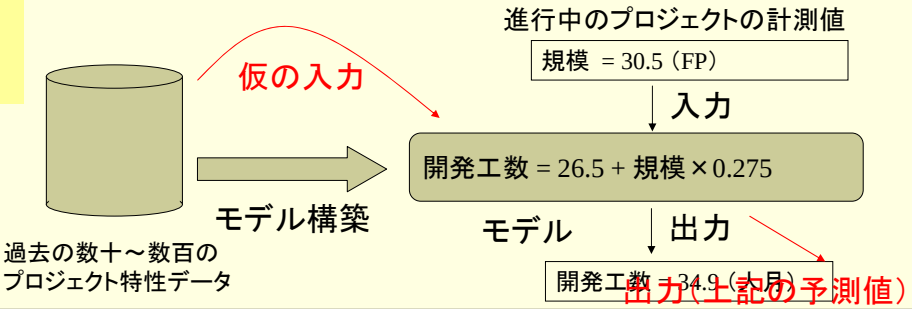
例題1の結果　注意点

予測結果

番号	観察値	予測値	残差	標準化残差
1	5152.000000	6992.955093	-1840.955093	-0.751814
2	5635.000000	5103.539044	531.460956	0.220733
3	805.000000	1715.085434	-910.085434	-0.375773
4	3829.000000	5092.929905	-1263.929905	-0.521641

. . . .

- この「予測値」は，「予測」により得られた値ではない！



例題1の結果

予測結果

番号	観察値	予測値	残差	標準化残差
1	5152.000000	6992.955093	-1840.955093	-0.751814
2	5635.000000	5103.539044	531.460956	0.220733
3	805.000000	1715.085434	-910.085434	-0.375773
4	3829.000000	5092.929905	-1263.929905	-0.521641

. . . .

$$\text{相対残差} = \frac{|\text{残差}|}{\text{実測値}}$$

相対残差の平均＝0.498

（精度はよくない）

例題2

- 例題1と同じデータセットを用い，目的変数，および，説明変数を対数変換してから重回帰モデルを作成せよ．（すなわち，対数線形モデルを作成せよ）
  - ただし，チーム経験年数（ ExpEquip）とプロジェクトマネージャ経験年数（ExpProjMan ）はゼロが含まれるため，対数変換できない．そこで，この2つの変数については，全て＋1してから対数変換せよ．
  - また，結果の評価に使う「観測値」，「予測値」，「残差」は，対数変換前の値を算出せよ．

例題2の結果

変数	偏回帰係数	標準誤差	t値	P値	標準化偏回帰係数
Var01	0.3849775	0.1163379	3.30913	0.00147	0.2886319
Var02	-0.07782952	0.1455016	0.53490	0.59439	-0.04487862
Var03	0.1502131	0.1414241	1.06215	0.29177	0.0896984
Var04	0.6866678	0.1122953	6.11484	0.00000	0.5186158
Var05	-0.8407526	0.1429008	5.88347	0.00000	-0.4341687
定数項	1.630145	0.2228920	7.31361	0.00000	

↓ 重回帰モデル式

log(ActualEffort) =

0.3849775*log(Duration) – 0.07782952*log(ExpEquip)
+ 0.1502131*log(ExpProjMan) + 0.6866678* log(Adj FPs)
– 0.8407526*log(Dev Env) + 1.630145

例題2の結果

- 対数変換前の値に戻してから評価する.

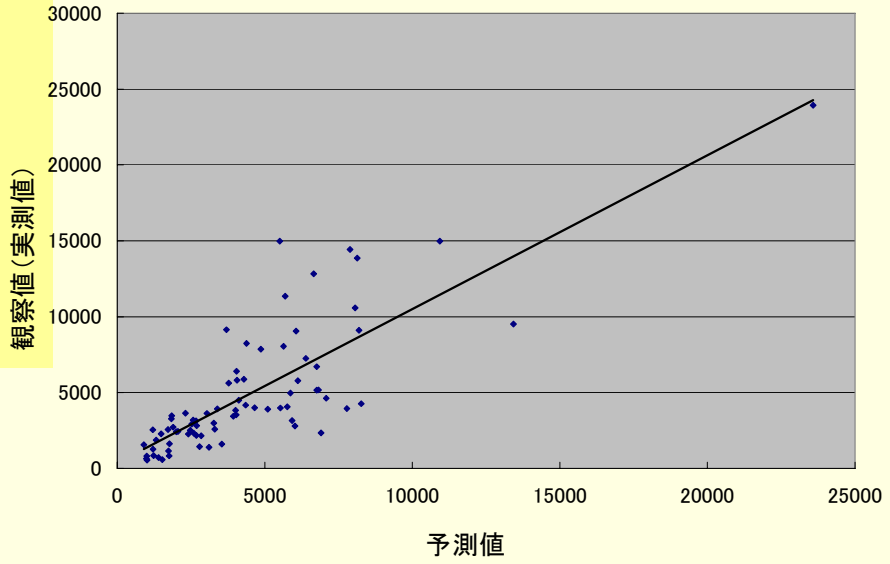
予測結果

番号	観察値	予測値	残差	標準化残差
1	3.711976	3.830112	-0.118136	-0.596390
2	3.750894	3.577434	0.173460	0.907528
3	2.905796	2.998508	-0.092712	-0.528460
. . . .				

$10^{観測値}$　$10^{予測値}$　$(10^{観測値} - 10^{予測値})$

番号	観察値	予測値	残差
1	5152.001728	6762.573526	-1610.571798
2	5635.001033	3779.496957	1855.504076
3	805.0002217	996.5704375	-191.5702157
…			

例題2の結果



例題2の結果

番号	観察値	予測値	残差
1	5152.001728	6762.573526	-1610.571798
2	5635.001033	3779.496957	1855.504076
3	805.0002217	996.5704375	-191.5702157
…			

相対残差の平均＝0.409

（例題1と比べて少し精度が向上した）

例題3

- canada.csvの77プロジェクトを，1986年以前に完了したプロジェクトと，1987年以降に完了したプロジェクトに2分せよ．（変数「Year Fin」を参照せよ）
  - 前者を，フィットデータセット（58プロジェクト）
  - 後者を，テストデータセット（19プロジェクト）と呼ぶことにする．
- フィットデータセットを用いて，例題2と同じ方法で対数線形モデルを作成せよ．
- 作成したモデルをテストデータに当てはめて，予測性能を評価せよ．（相対誤差の平均値を求めよ）

より現実的な予測

- 問い：project1.csvで同様のことができるだろうか？
- 回答：そのままではできない
  - 欠損値が存在する．
  - カテゴリ尺度が存在する．
  - 説明変数の候補が多すぎる．

より現実的な予測の手順

1. 予測を行う開発工程の決定
   - 例：詳細設計完了時
   - 説明変数の候補が決まる.
2. 欠損値のないデータセットの作成
   - いくつかのプロジェクト，変数の削除
   - 欠損値の補完
3. 尺度の変換
   - カテゴリ変数→数値変数
4. モデルの構築
5. モデルの評価

予測を行う開発工程の決定

- 工程の例：
  - システム化計画完了時
  - 要件定義完了時
  - 基本設計完了時
  - 詳細設計完了時
  - コーディング完了時

システム化計画完了時

- システム化計画　実績工数（人時）
- 開発プロジェクト種別
- 母体システム安定度
- 開発プロジェクト形態
- 新規顧客
- 新規業種・業務
- 新規協力会社
- 役割分担 責任所在
- 業種
- 業務種類
- 開発ライフサイクルモデル
- 類似プロジェクトの有無
- ユーザ担当者 システム経験
- 開発期間（月数）計画値
- PM（プロジェクトマネージャ）スキル

要件定義完了時

- システム化計画完了時に利用可能な変数
- 利用形態
- 利用拠点数
- システム種別
- 処理形態
- 要求仕様_明確度合
- ユーザ担当者_要求仕様関与
- 要求レベル_信頼性
- 要求レベル_使用性
- 要求レベル_性能・効率性
- 要求レベル_保守性
- 要求レベル_移植性
- 要求レベル_ランニングコスト要求
- 要求レベル_セキュリティ
- 業務パッケージ_利用有無
- 達成目標_優先度_明確度合
- 法的規制有無
- 要件定義書 文書量
- 月数（実績）要件定義
- 要件定義　実績工数（人時）

基本設計完了時

- 要件定義完了時に利用可能な変数
- 主開発言語
- 開発言語数
- DBMSの利用
- プロジェクト管理ツール_利用
- 設計支援ツール利用
- ドキュメント作成ツール利用
- デバッグ_テストツール利用
- 上流CASEツール利用
- コードジェネレータ利用
- 主なFP計測手法
- FP実測値_調整前
- ILF実績値
- EIF実績
- トランザクションファンクション実績値
- データファンクション実績値
- 設計書文書量基本設計書
- 月数(実績)基本設計
- 基本設計書レビュー指摘件数
- 基本設計　実績工数(人時)
- アーキテクチャ
- アーキテクチャ数
- 開発対象プラットフォーム
- 開発対象プラットフォーム数
- Web技術の利用
- 開発方法論利用

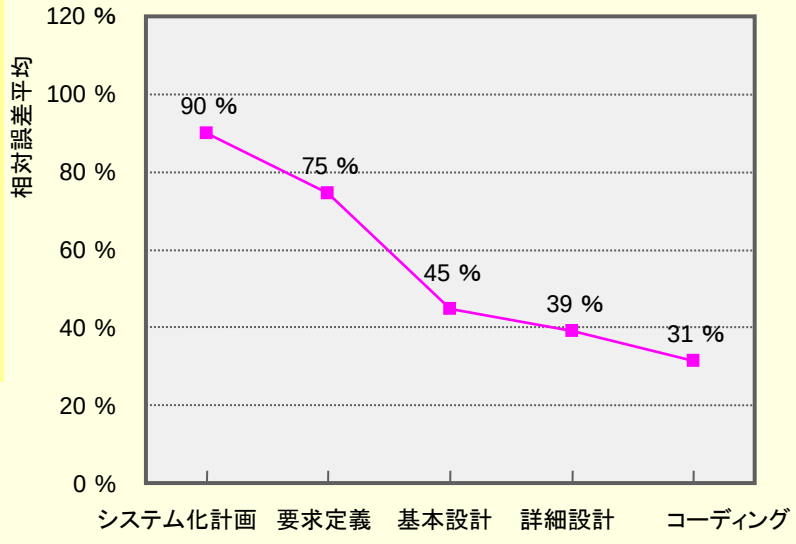
詳細設計完了時

- 基本設計完了時に利用可能な変数
- 構成管理ツール利用
- ユーザ担当者_設計内容理解度
- 要員スキル_業務分野経験
- 要員スキル_分析・設計経験
- 要員スキル_言語・ツール利用経験
- 要員スキル_開発プラットフォーム使用経験
- 設計書文書量詳細設計書
- 規模指標_DBテーブル数
- 規模指標_画面数
- 規模指標_帳票数
- 規模指標_バッチ本数
- 月数(実績)詳細設計
- 詳細設計書レビュー指摘件数
- 詳細設計　実績工数(人時)

コーディング完了時

- 詳細設計完了時に利用可能な変数
- ソースコード再利用率
- SLOC実測値
- 月数(実績)製作
- 外注実績(金額比率)
- コーディング 実績工数(人時)
- 外部委託率

予測の時期と誤差の関係(例)



欠損値のないデータセットの作成

- 欠損値処理法
  - 平均値挿入法：欠損値に対して当該変数の平均値を挿入する．
  - リストワイズ除去法：欠損値を一つでも含むケースを削除する．
  - 他に，ペアワイズ除去法，ホットデック法，k-NN法などがある．

- 現実的には，

【 手順0 】必要不可欠な変数が欠損しているプロジェクトを削除する．
例：規模（FP）が欠損しているプロジェクト
【 手順1 】欠損率の高い変数を除去する（30%以上）．
【 手順2 】欠損率の高いプロジェクトを除去する（30%以上）．
【 手順3 】欠損値を補完する．
カテゴリ変数については，「その他」という値（カテゴリ）を設ける．

変数の変換

プロジェクトID	業種
PRO-01	銀行
PRO-02	製造業
PRO-03	銀行
PRO-04	銀行
PRO-05	製造業
PRO-06	銀行
PRO-07	銀行
PRO-08	公共

カテゴリ変数
（名義尺度）

ダミー変数化
（2値化）

業種＝銀行	業種＝製造業	業種＝公共
1	0	0
0	1	0
1	0	0
1	0	0
0	1	0
1	0	0
1	0	0
0	0	1

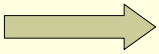
ダミー変数

便宜上，量的データ（間隔尺度，比尺度）とみなす．

変数の変換2

プロジェクトID	要求仕様明確度合い
PRO-01	やや明確
PRO-02	非常に明確
PRO-03	やや曖昧
PRO-04	やや曖昧
PRO-05	やや明確
PRO-06	非常に曖昧
PRO-07	やや曖昧
PRO-08	非常に明確

順序尺度



変数化

要求仕様明確度合い
3
4
2
2
3
1
2
4

便宜上，量的データ（間隔尺度，比尺度）とみなす．

例題4

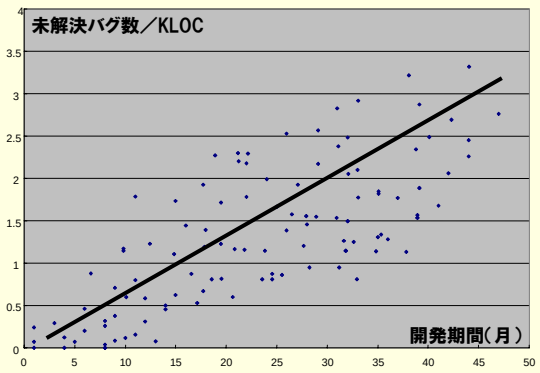
- project1.csvから欠損値のないデータセットを作成せよ．
- 詳細設計完了時を想定し，開発工数を予測する対数線形モデルを作成せよ．ただし，
  - 「開発期間」「ピーク要員数」は，計画値とみなしてよい．（説明変数として用いてよい）

プロジェクト特性データに基づく見積もり

- 定義済みモデルに基づく見積もり
  - COCOMO，COCOMO II，Agile COCOMO
- 過去のデータに基づく見積もり
  - モデルベース手法
    - 重回帰分析，CoBRA法，ニューラルネット，．．．
  - メモリベース手法
    - Analogy-based法，協調フィルタリング法，OSR法

モデルベース手法の問題点（1）

- 多様なソフトウェア開発プロジェクトを一つのモデルで表現することは難しい．



モデルへの不信

- モデルによる見積もり値をどこまで信用してよいか分からない.
  - プロジェクトは個別性の高いものである.
  - あるプロジェクトに当てはまるからといって，他のプロジェクトにも当てはまるとは限らない.

モデルベース手法の問題点（2）

- データ欠損に対して脆弱である.
  - データ欠損を補う方法は開発されているが，欠損率が30%を超えると，予測精度は著しく低下する.

Kromrey, J., and Hines, C.: “Non-randomly missing data in multiple regression: An empirical comparison of common missing-data treatments,” *Educational and Psychological Measurement*, 54, 3, pp.573-593 (1994).

データの欠損は避けられない

- 開発過程のデータ（リアルタイムに収集されるデータ）は，取り直しがきかない．

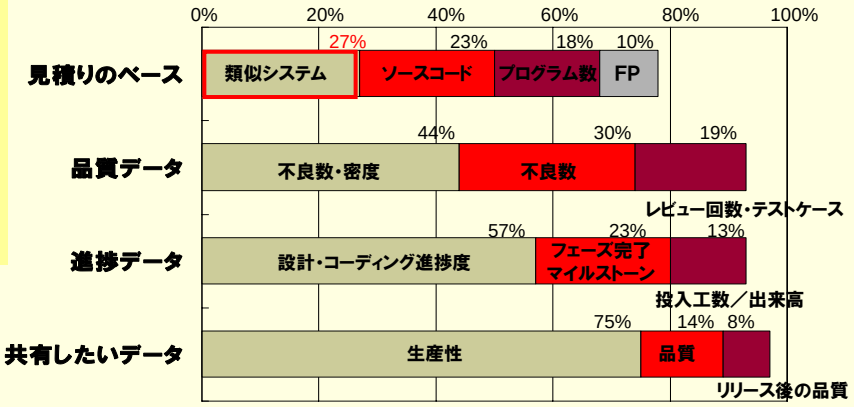
		項目 1	項目 2	項目 3	・・・	項目 490
A 社	A社プロジェクト 1	値 1-1	値 1-2	欠損値	・・・	欠損値
	A社プロジェクト 2	値 2-1	値 2-2	欠損値	・・・	欠損値
	A社プロジェクト 3	値 3-1	値 3-2	欠損値	・・・	欠損値
B 社	B社プロジェクト 4	欠損値	値 4-2	値 4-3	・・・	欠損値
	B社プロジェクト 5	欠損値	値 5-2	値 5-3	・・・	欠損値
	B社プロジェクト 6	欠損値	値 6-2	値 6-3	・・・	欠損値
C 社	C社プロジェクト 7	欠損値	欠損値	値 7-3	・・・	値 7-490
	C社プロジェクト 8	欠損値	欠損値	値 8-3	・・・	値 8-490
	C社プロジェクト 9	欠損値	欠損値	値 9-3	・・・	値 9-490

モデル構築より類似プロジェクト検索

- 有能なプロジェクト管理者は，見積りや問題解決において，オールマイティなモデルを持っているわけではない．
- モデルよりも個々のデータを経験としてうまく活用している．
  - 過去に携わったプロジェクト群の中から，現行プロジェクトと似たプロジェクト（類似プロジェクト）を選び出す．
  - 類似プロジェクトにおける開発コスト，作業進捗，発生した問題とその解決策，などを，現行プロジェクトに（多少アレンジした上で）適用する．

受注ソフトウェア開発での計測

（社）情報サービス産業協会（JISA），「情報サービス産業における受注ソフトウェア開発の技術課題に関わるアンケート調査」，2004年．

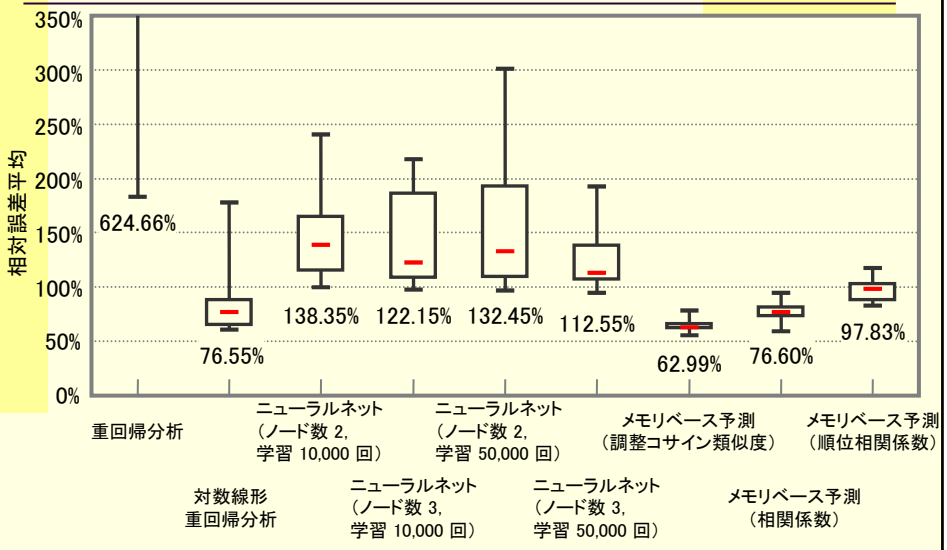


メモリベース予測

- ステップ1: 類似度計算
  - 説明変数の値に基づいて，現行プロジェクトXと過去プロジェクトそれぞれ（A，B，C，…）との間で類似度を計算する．
- ステップ2: 予測値計算
  - 現行プロジェクトXと類似度の高い *k*個の過去プロジェクトの工数を類似度で加重平均して，現行プロジェクトXの工数の予測値とする．

	設計工数	製造工数	基本設計欠陥数	詳細設計欠陥数	予測結果 40.0
現行プロジェクトX	50	20	3	10	
過去プロジェクトA（類似度: 0.99）	45	18	2	(欠損値)	36
過去プロジェクトB（類似度: 0.99）	(欠損値)	22	3	11	44
過去プロジェクトC（類似度: 0.69）	10	10	(欠損値)	5	30

メモリベース予測の評価事例



出典：大杉 他, “企業横断的収集データに基づくソフトウェア開発プロジェクトの工数見積もり,” SEC journal, No.5, pp.16-25, February 2006.

分析の目的

- データ間の関係を調べる.
  - 視覚的に
    - 散布図，ヒストグラム，箱ひげ図，平行座標プロットなど
  - 定量的に
    - t検定，カイ二乗検定，分散分析，無相関検定など
    - 相関係数，クラメールのV，回帰曲線など
- データの予測（見積もり）を行う.
  - 重回帰分析，協調フィルタリング，マハラノビスタグチ法など
- 大量のデータの中から隠された関係を発見する.
  - アソシエーション分析（相関ルール分析）

アソシエーション分析（相関ルール分析）

- 事象間の強い関係（アソシエーションルール）を発見する手法である.
- アソシエーションルール：A⇒B
  - ある事象Aが発生するならばある事象Bも高い確率で発生する.
- コンビニの購買履歴から得たアソシエーションルールの例
  - 休日に「レジャーシート」を買う顧客は「おにぎり」と「お茶」も同時に買っている。
    「（曜日＝土日） and おにぎり and お茶 ⇒ レジャーシート」

→ 休日には、レジャーシートの配置をおにぎりかお茶に近づけ、発見率、併せ買い率を上げる。

アソシエーション分析（相関ルール分析）

- プロジェクト特性データの場合
  - 「（開発種別=拡張） and （アーキテクチャ=3階層CS）⇒テスト工数比率＝大」
    3階層アーキテクチャの機能拡張プロジェクトではテスト工数比率が高くなる.

→ 3階層アーキテクチャの機能拡張プロジェクトのテスト工数は他よりも大きく見積る必要がある.

アソシエーションルールに関する指標

- アソシエーションルール $X \Rightarrow Y$ に対して，

$$\text{支持度} = \frac{X\text{と}Y\text{が同時に出現したケース数}}{\text{全ケース数}}$$

支持度が大きいほど，よく起こる事象を表すルールであると言える．

$$\text{信頼度} = \frac{X\text{の発生時に}Y\text{が出現したケース数}}{X\text{が発生したケース数}}$$

$$\text{リフト値} = \frac{X \Rightarrow Y\text{の信頼度}}{\text{全ケースにおける}Y\text{の発生割合}}$$

信頼度とリフト値は，前提Xと結論Yの関連の強さを表す指標である．

アソシエーションルール適用の前提

- 全ての変数は，カテゴリ変数である．
  - 数値変数はあらかじめカテゴリ変数に変換しておく．
  - 欠損値があってもよい．

分析データのイメージ

業種	アーキテクチャ	主開発言語	規模（FP）	工期（月数）	平均要員数	外部委託率	生産性（FP÷人時）
銀行	C/S	PL/I	上位	上位	上位	上位	下位
	スタンドアロン	C	下位	中位	中位	下位	下位
銀行	C/S	COBOL	下位	中位	下位	下位	下位
銀行		Visual Basic	中位		中位	中位	中位
製造業	スタンドアロン		中位	下位	中位	中位	中位
銀行	混合	C++		下位	上位	上位	上位
銀行	C/S	COBOL	上位	下位	下位	下位	上位
公共	混合	Java	中位	上位		下位	上位

抽出したルールの例

アーキテクチャ=スタンドアロン 100
⇒ 外注率=ゼロ 85　conf:(0.85)

スタンドアロンシステムの開発は，外注しないことが多い．

PMスキル=小・中規模プロジェクトの管理しか経験していない 103
⇒ 外注率=ゼロ 84　conf:(0.82)

スキルの低いPMを使うときは，外注しないことが多い．

主開発言語=VB 130
⇒ 主開発プラットフォーム=Windows系 99　conf:(0.76)

言語がVBであるならば，プラットフォームはWindows系である．

例題5

- canada.csvもしくはproject1.csvから
  - 支持度（support）0.1以上
  - 信頼度（confidence）0.8以上

  のアソシエーションルールを抽出せよ．また，
  - 支持度（support）0.1以上
  - リフト値（lift）2.0以上

  のアソシエーションルールを抽出せよ．
- 抽出したルールのうち，結論部が「生産性＝低い」もしくは「生産性＝高い」となるものを抜き出せ．
- アソシエーションルール抽出ツールとして，WEKAが利用できる．

演習課題（前回出題）をお忘れなく！

- レポート提出期限
  - 2006年12月22日（金）2限
  - 希望者はレポート課題の内容を発表すること
    - レポートの内容について説明する．
    - レポート用紙をスクリーンに映す，もしくは，パワーポイント等のプレゼン資料を用いる．
  - 時間が余れば，提出されたレポートの中からいくつかを講義中に（門田が）紹介します．

- 連絡先
  - 門田暁人 akito-m@is.naist.jp
  - B303室，内線5311

今回の講義で紹介した分析技術を用いてもよい．