

# 社会情報データを利用した統計教育

## —解析事例データベースの構築へ向けて—

森 裕一・太田 明宏\*

岡山理科大学総合情報学部社会情報学科

\*岡山理科大学大学院総合情報研究科社会情報専攻

(2008年9月30日受付、2008年11月7日受理)

### 1. はじめに

経済、経営、消費、人口、世論調査など、社会情報のデータには、さまざまなものがある。これらのデータは、統計教育、特に社会系のデータ解析の学習の生きた題材となる。最近では、社会科学データベースなど、インターネットを中心として、データのアーカイブが構築され、統計教育にも利用しやすくなっている。

一方、統計教育に目を向けると、実際の(生の)データを利用することは大変重要であるが、データだけがあっても教育にはならず、そのデータをどのように料理していくかがしっかりと身につかなければ、実践で役に立たない。すなわち、解析目的に到達するために正しく解析手段を選び、正しく結果を解釈できる力をつけ、同時に、各種の統計専用パッケージを使いこなせるようにすることが重要である。

これらに対して、Mori et al. (2003) は、Data oriented statistical system (データ指向統計解析環境, **DoSS@d**) を提唱し、システムを <http://mo161.soci.ous.ac.jp/@d/> に構築・稼働させている。これは、データとそのデータの解析事例をデータベース化したもので、データとともに、そのデータをどのように解析したかの手順を示したドキュメント(解析ストーリーとよぶ)も一緒にアーカイブ化したものである。この解析ストーリーもデータベース化しているのは、過去の解析事例としての記録の意味だけでなく、将来同様の手順を行う際の見本としても利用できるようにするためである。また、その解析ストーリーと同じ結果を、SPSS, JMP, R, XploRe, Excel といった統計ソフトでどのように得るかの解説もつけており(一部オンラインでも解析が可能となっている)、解析ストーリーを核とした統計教育が非常にやりやすくなっている。

このように、さまざまなデータを解析ストーリーに従って学習していく環境が、**DoSS@d** によって提供されているので、正しい手法の選択、それに応じた統計ソフトの実行、正しい結果の解釈に対する力をつける統計教育が容易に展開できることわかる。しかし、現在の **DoSS@d** に蓄えられている解析ストーリーは、体系的に整理されておらず、いくつか用意されている解析ストーリーに基づく学習コースも試験的であるので、多くの事例の収集・蓄積とともに、体系的な分類およびその分類に基づく教育コースの提供が望まれるところとなっている。

そこで、本研究では、主として社会情報データのデータセットと解析ストーリーを数多く集め、それを分類・整理することで、**DoSS@d** の事例データベースの充実を図ることとする。

### 2. DoSS@d について

**DoSS@d** では、次の作業や機能・役割の実装を目的とされている(森 他, 2007)。

- 自由に使える実データセットを数多く収集し、データベース化すること。
- 解析ストーリーを数多く収集し、データベース化すること。
- データセットや解析ストーリーでは、その目的、解析分野、出典などを明確に示すこと。
- c)の各内容は決められた様式で記述されること。
- 解析ストーリーについては、そこに記述されている解析手順を一般の統計パッケージを用いてどのようにして実現するかの手順書もつけ、解析ストーリーの再現に供すること。
- データベース内のデータセットや解析ストーリーを分野、手法などの項目内の検索語によって検索で

きる検索機能をもつこと。

- g) データベース内の任意のデータセットをその場で解析できるように、オンライン解析機能をもつこと。
- h) 解析ストーリーをもつデータセットをオンライン解析機能で解析する際、解析ストーリーどおりに解析できるようにパラメータなどが自動設定される仕組みをもつこと。
- i) 複数の解析ストーリーを教育目的に応じて配列した学習コースを提供すること。
- j) 用語集や解説のページを用意すること。
- k) 新たにデータを収集したり、それらをデータベースに登録したりできるシステムを提供すること。
- l) データや解析ストーリーの評価を行う機能をもつこと。
- m) 多言語へ対応できること。

現在、上記の各機能を実現させるため、**DoSS@d** は、次の3つのサブシステムをもっている。

・ **DoDStat@d** (データ指向統計データベース, Data oriented Database of Statistics)

これは、実際のデータセットと解析ストーリーのデータベースである。各データは、CSV フォーマットのデータファイル、属性情報(ケース名、変数名および変数の型など)を含んでおり、検索エンジンにより対象分野や解析手法、キーワードといった検索キーを使用して、目的に応じたデータセットを選択することができるようになっている。また、解析ストーリーにおいては、必要に応じて、そのストーリーに基づいた解析がオンラインで行えるようになっている(実際には、次の **DoAStat@d** へのパラメータ付きリンクである)。

・ **DoAStat@d** (データ指向統計解析システム, Data oriented Analysis system of Statistics)

これは、サーバ側の統計エンジン(RとXploRe Quantlet Server)を利用した解析環境を提供する。ユーザは、**DoDStat@d** に格納されたすべてのデータを解析でき、かつ、ユーザ保有のデータも解析できる(詳細は、Honda et al., 2004などを参照のこと)。

・ **DoLStat@d** (データ指向統計学習システム, Data oriented Learning system of Statistics)

これは、解析ストーリーに基づいて、統計を学習できるようにした学習コースを Web 上で提供するものである。具体的には、**DoDStat@d** が保有する解析ストーリーを学習目標に合わせて教育的に配列したもので、個々の解析ストーリーの学習時には、**DoAStat@d** のオンライン解析も利用できる。本研究で追加していく解析ストーリーは、この **DoLStat@d** で、教育コースのコンテンツとして利用することになる。

これら3つのサブシステムに加え、k)のデータ収集・登録機能が、モバイル統計の可能性も含めて検討され(Mori et al., 2006; 森 他, 2006, 2007), l)のデータや解析ストーリーの評価を行う機能についても、プロトタイプとして実装が試みられている(森 他, 2008)。また、**DoLStat@d** の学習システムを拡張する機能の開発も行われている(山崎, 2005; 宿久 他, 2008)。

今回は、上記の目指す機能のうち、b), c), e), 特に、i)を充実させることが目的である。

### 3. 解析事例のストーリー化

#### 3.1 事例の収集

現在の **DoSS@d** には、18種類の解析ストーリーが蓄積されている(表2)。

これに対して、社会情報データとして、「経済」分野のデータを中心に、その解析ストーリーを増やすため、著者の許可を得て、次の4冊の事例集にある事例のストーリー化を行うことにした。

- データマイニング事例集 [38] (上田, 1998)
- データマイニング実践集 [37] (上田, 1999)
- 実践ワークショップ Excel 徹底活用 多変量解析 [30] (上田 他, 2003)
- 実践ワークショップ Excel 徹底活用 ビジネスデータ分析 [18] (淵上 他, 2004)

いずれの書籍にも経済やマーケティングに関するデータを例に、統計の基礎的・一般的な手法を用いて、実データ解析に役立つ事例が掲載されている(表4~7を参照)。上記の[ ]内の数値は、各書籍で取り上げられている解析事例の数である。4冊で、合計123種類の事例がある。これらを次に示す方法で分類し、解析ストーリーに変換していくこととした。

3.2 解析事例の分類と解析ストーリーへの変換

DoLStat@d では、学習コースとして、

統計学一般コース 専門分野別コース 統計技法学習コース 手法別学習コース  
 の4つの大きなカテゴリー（コース分類）を設け、その各カテゴリー内で複数のテーマ（収録コース）を設定している（表1）。すなわち、収録された解析ストーリーは、このカテゴリーのどれにあたるか分類・整理され、各収録コースは、該当する解析ストーリーの中から2~7種を選んで構成されている。表2は、DoDStat@d に保有されている解析ストーリー18種がどの分類にあてはまり（●や○）、具体的に、どの学習コースで使われているか（●）を示したものである。1つの解析ストーリーに該当する収録コースは1つ以上で、かつ1つの学習コースでは複数の解析ストーリーが利用できることがわかる。

表1 DoLStat@d の学習コース

コース分類	収録コース
統計学一般コース	統計学入門コース1・2 統計学一般コース1・2 多変量解析入門コース 応用統計学コース
専門分野別コース	マーケティングコース 生活と統計コース 経済学コース 生物統計コース
統計技法学習コース	要約コース 視覚化コース 比較と検定コース 予測コース 分類コース 次元縮約コース
手法別学習コース	記述統計コース 回帰分析コース 主成分分析・コレスポネンス分析コース クラスター分析と判別分析コース 因子分析コース

表2 既存の解析ストーリーの収録状況

解析ストーリー	分野		専門分野				統計技法					手法					
	大分類	小分類	マーケティング	生活と統計	経済学	生物統計	要約	視覚化	比較と検定	予測	分類	次元縮約	記述統計	回帰分析	PCA・CA	CLS・DISC	因子分析
タイタニック号で生死を分けた要因	社	社					●	●					●				
タイタニック号データの独立性検定	社	社							●								
最高経営責任者の給料と年齢	経	金			○								●				
スイス銀行紙幣の真札と偽札の違い	社	社					●	●					●				
スイス銀行紙幣の真札と偽札の検定	社	社							●				●				
缶コーヒーの好感度（相関）	経	消	○	○	○		●	○					●				
缶コーヒーの好感度	経	消	●	○	○				●					●			
新店舗の売上予測	経	経	●		○					●			○	●			
セメントの発熱量の予測	工	工							●					●			
売れ筋商品と在庫管理	経	消	●		○									●			
スーパーの客数予測	経	経	●		○					○							
世界の都市の経済 1	経	経	●		○						○						
世界の都市の経済 2	経	経	●		○						○					○	
都道府県別アルコール類の消費量	経	消	●	○	○												
羽根アリの身体的特徴	科	生				○										●	
麺つゆの特徴	経	消	●	○	○			●								●	
家庭用 PC サーバの特徴	経	消	○	○	○											●	
ペットフードの特徴	経	消	○	○	○											●	

PCA：主成分分析，CA：コレスポネンス分析，CLS：クラスター分析，DISC：判別分析  
 ●：収録されている学習コース  
 ○：収録可能な学習コース

そこで、3.1の4冊に掲載されている123の解析事例すべてに対して、上記のカテゴリーのどれにあたるかを整理することにする。その結果をもとに、解析ストーリーへの変換を順に施していく。

全123事例の分類・整理の結果を付録の表4~7に示す。

今回の分類とストーリー化の作業により、経済分野90種、社会科学分野18種、工業分野8種、スポーツ分野4種、医療分野1種、科学分野1種、その他1種の解析ストーリーが学習コースの単元として追加できるようになり、既存の学習コースに収録可能となる解析ストーリーの数は、表3のようになる。因子分析については、その手法を利用する解析ストーリーはなかった。

なお、対象とした4冊の事例集では、ABC分析、フェイスグラフ、ダミー変数による回帰分析、判別分析、数量化I類、数量化II類、数量化III類、双対尺度法（多次元尺度構成法）、コンジョイント分析などの各手法が紹介・適用されているが、表2~7では、各解析ストーリーを現在のDoLStat@dに収録されているコース分類で整理をした（多くのものが回帰分析やコレスポネンス分析に分類されている）ので、それらの手法名は表に出ていない。しかし、今回の作業で、これら多くの手法が追加されたことで、それぞれの手法に焦点をあてた分類を新たに設けることにより、より多様な学習コースの構築も可能となっている。

表3 DoLStat@dの学習コースへの収録可能数

コース分類	収録コース	収録可能数
専門分野別コース	マーケティングコース	72
	生活と統計コース	57
	経済学コース	87
	生物統計コース	1
統計技法学習コース	要約コース	6
	視覚化コース	91
	比較と検定コース	15
	予測コース	58
	分類コース	6
	次元縮約コース	21
手法別学習コース	記述統計コース	28
	回帰分析コース	76
	主成分分析・コレスポネンス分析コース	22
	クラスター分析と判別分析コース	3
	因子分析コース	0

#### 4. 事例

作成された解析ストーリーのページでは、次のことが可能となっている。

- そのデータがどんな目的で、どの手法を使って解析され、その結果がどのように解釈されたかの確認
- 元データの表示
- 解析ストーリーに基づくオンライン解析の実行
- 統計パッケージ（SPSS, JMP, R, XploRe, Excel）で、その解析ストーリーを実現するにはどのようにすればよいかの解説の表示

3で作成した解析ストーリーのうち、「アパートの専有面積と賃料の関係」（表7）と「DM（ダイレクトメール）のレスポンスの判別予測」（表6）を例に、具体的な内容を示す。

##### (1) 「アパートの専有面積と賃料の関係」のデータセットと解析ストーリー

図1が「アパートの専有面積と賃料の関係」の解析ストーリーのページである。解析ストーリーの「タイトル」、「ゴール」、「データ」（元のデータへのリンク）、「出典」、「分野」が示され、その下に、具体的な解析手順である「解析ストーリー」が示されている（このフォーマットは、すべての解析ストーリーページで共通である）。ここでは、賃料を $y$ 、専有面積を $x$ としたときの散布図を作画・考察することで、「専有面積が広いほど、賃料が高くなること」が結論として得られるという比較的単純なものである。右下の[Excel]ボタンをクリックすると、この解析をExcelで行うときの操作説明が表示される（具体例は、事例(2)で示す）。また、元データへのリンク[アパート物件データ1]をクリックすると、図2のデータページが開く。

##### (2) 「DM（ダイレクトメール）のレスポンスの判別予測」のデータと解析ストーリー

図3が「DMのレスポンスの判別予測」の解析ストーリーで、まず、全変数を用いて回帰分析を行い、そ

の結果のあてはまりがよくないため、1つの変数を削除してから再び回帰分析を行う、というやや複雑な解析手順をとる事例となっている。元データへのリンク [DM レスポンスの有無のデータ] をクリックすると、データページが開く (スクリーンショットは省略)。一方、このページの右下には、[SPSS], [JMP], [R], [XploRe], [Excel] のボタンが見える。それぞれのボタンをクリックすれば、今回の解析ストーリーを実現するための各統計解析パッケージの操作説明が表示される。図 4 に、Excel の場合の操作説明ページを示す。(左下の [Analysis] ボタンの機能については、Honda et al. (2004) や森 他 (2007) を参照のこと)

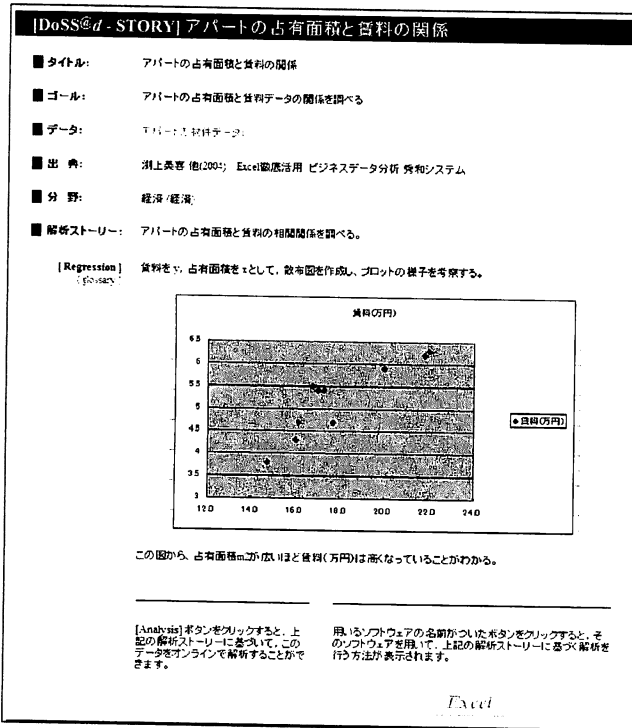


図 1 「アパートの専有面積と賃料の関係」の解析ストーリーページ

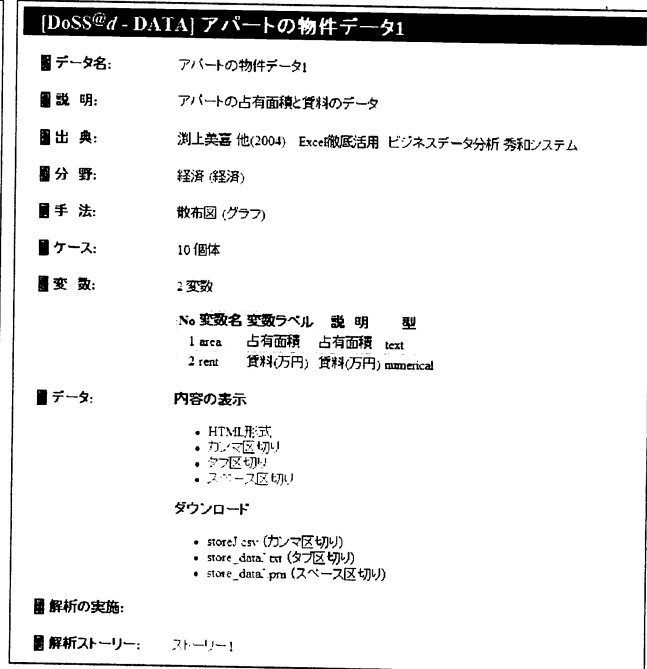


図 2 「アパートの専有面積と賃料の関係」のデータページ

### 5. おわりに

本研究では、社会情報データのデータセットと解析ストーリーを数多く集め、データ指向統計解析環境 **DoSS@d** の事例データベースの充実を図ることを目指した。そのために、市販の事例集から、実データと解析ストーリーを取り出し、用いられている分野や手法を洗い出し、**DoSS@d** の分類にあてはめ、整理し、さらに、各事例を **DoSS@d** の仕様へ変換することを行った。これにより、これまで 18 種類であった解析ストーリーに加え、今回対象とした 123 種類の解析ストーリーを利用することで、各種の学習コースの作成が可能となるストーリー数をもつデータベースへと拡張することができたといえる。このことは、さまざまなデータを解析ストーリーに従って解析していくことにより、正しい手法を選択し、それに応じて統計ソフトを操作し、そして正しく結果を解釈する力をつけるための環境の提供に大きく貢献するものと考えられる。

今後の課題としては、作成した解析ストーリーを利用した学習コースを、さまざまなユーザの要求に耐えられるように提供していくことと、作成した解析ストーリーの多言語化、および数多くのデータセットや解析ストーリーを多角的に統計学習に利用できる教育システムを提供していくことなどがあげられる。

### 参考文献

Honda, K., Mori, Y., Yamamoto, Y. and Yadohisa, H. (2004) Web-Based Analysis System in Data-oriented Statistical System "DoSS@d". In: Antoch, J (ed), *COMPSTAT2004 Proceedings in Computational Statistics*, 1209-1216, Physica-Verlag.

Mori, Y., Yamamoto, Y. and Yadohisa, H. (2003) Data-oriented Learning System of Statistics based on Analysis Scenario/Story (DoLStat). *Bulletin of the International Statistical Institute, 54th Session Invited Papers*, Volume LX Two Books, Book 2, 74-77.

Mori, Y., Yamamoto, Y. and Yadohisa, H. (2006) Data collection and document generation system for data-oriented approaches. The 17th Symposium on Computational Statistics, 218, Abstract Book; 1633-1640, Proceedings CD.

**[DoSS@d-STORY] DMのレスポンスの有無を判別予測する**

■ **タイトル:** DMのレスポンスの有無を判別予測する

■ **ゴール:** 居住地、年収、年齢で、DM(ダイレクトメール)へのレスポンスの有無をまとめたデータから、DMのレスポンスの有無を判別予測する最適式を求める。

■ **データ:** DMレスポンスの有無のデータ

■ **出典:** 実践ワークブック Excel徹底活用 多変量解析 上田太一郎 菊田正雄 本田和寛

■ **分野:** 経済(経済)

■ **解析ストーリー:** 居住地、年収、年齢で、DM(ダイレクトメール)へのレスポンスの有無をまとめたデータから、DMのレスポンスの有無を判別予測する最適式を求め、No.53を除いたデータに対してデータを予測する。

[Regression] 「DMレスポンスの有無」を y、残りの変数を x として、回帰分析を行う。

概要

回帰統計	
重相関 R	0.594714
重決定 R2	0.353664
補正 R2	0.31329
標準誤差	0.40709
観測数	52

分散分析表					
	自由度	変動	分散	観測された分散比	有意 F
回帰	3	4.353039	1.451013	8.755708888	9.76E-05
残差	48	7.954653	0.165722		
合計	51	12.30769			

	係数	標準誤差	t	P-値
切片	-0.96164	0.683476	-1.3867	0.171940627
都心	0.42869	0.120016	3.580266	0.000798069
年収(単位:10万円)	0.009608	0.002767	3.471892	0.001103836
年齢	-0.00045	0.009647	-0.04687	0.962814268

年齢のP-値(危険率)が一番大きいので、これを削除し、あらためて回帰分析を行う。実行結果は次回の上図になる。

概要

回帰統計	
重相関 R	0.594689
重決定 R2	0.353655
補正 R2	0.327273
標準誤差	0.402923
観測数	52

分散分析表					
	自由度	変動	分散	観測された分散比	有意 F
回帰	2	4.352675	2.176338	13.40544471	2.27185E-05
残差	49	7.955017	0.162347		
合計	51	12.30769			

	係数	標準誤差	t	P-値
切片	-0.99009	0.332037	-2.98186	0.00445355
都心	0.43141	0.113088	3.814488	0.000382858
年収(単位:10万円)	0.009613	0.002737	3.512126	0.000655585

となり、回帰式は、次のようになる。

$$y = -0.90 + 0.431 \times \text{都心} + 0.010 \times \text{年収(単位:10万円)}$$

これに対して、No.53の各値(都心=1, 年収=10)を上記の回帰式に代入すると、「DMレスポンスの有無」の値は0.46となり、0.5より少ないので、レスポンスなし(=0)と予測される。実際の値は0なので、予想は的中しているといえる。

[Analysis] ボタン 用いるソフトウェアの名前がボタンをクリックすると、そのソフトウェアを用いて、上記の解析ストーリーに基づいて、このデータをグラフで解析することが出来ます。

Analysis SPSS JMF R NploRe Excel

図3 「DMのレスポンスの判別予測」の解析ストーリーページ

**Excelによる解析**

■ **DMのレスポンス有無を判別予測**

データ: [1.F] 1

0-1) 目標の確認  
居住地、年収、年齢で、DM(ダイレクトメール)へのレスポンスの有無をまとめたデータから、DMのレスポンスの有無を判別予測する最適式を求め、No.53のデータを予測する。

0-2) 解析の準備  
データを開く、「(ファイル)開く」

A	B	C	D	E
NO	都心	年収(単位:10万円)	年齢	DMレスポンスの有無
1	0	100	58	0
2	1	89	60	0
3	2	86	65	0
4	3	112	60	0
5	4	106	50	0
6	6	105	49	0
7	6	112	61	0
8	0	121	58	0
9	0	117	51	0
10	8	117	61	1
11	10	87	67	0
12	11	89	51	0
13	12	100	58	0
14	13	128	60	0
15	14	139	52	0
16	15	139	52	0

(1) 回帰分析  
操作: 「ツール」メニューまたは「データタブ」(2007)「右端」から「分析ツール」を選択し「回帰分析」を選ぶ。

「DMレスポンスの有無」を y として、残りの変数を x として、回帰分析を実行する。(ラベル変数を含んでデータを指定し、「ラベル」をチェックする。)

出力:

概要

回帰統計	
重相関 R	0.594714
重決定 R2	0.353664
補正 R2	0.31329
標準誤差	0.40709
観測数	52

分散分析表					
	自由度	変動	分散	観測された分散比	有意 F
回帰	3	4.353039	1.451013	8.755708888	9.76E-05
残差	48	7.954653	0.165722		
合計	51	12.30769			

	係数	標準誤差	t	P-値
切片	-0.96164	0.683476	-1.3867	0.171940627
都心	0.42869	0.120016	3.580266	0.000798069
年収(単位:10万円)	0.009608	0.002767	3.471892	0.001103836
年齢	-0.00045	0.009647	-0.04687	0.962814268

年齢のP-値(危険率)が一番大きいのでこれを削除して回帰分析を実行する。実行結果は下図のようになる。

A	B	C	D	E	F	G	H	I	
1	回帰								
2	回帰統計								
3	重相関 R	0.594689							
4	重決定 R2	0.353655							
5	補正 R2	0.327273							
6	標準誤差	0.402923							
7	観測数	52							
8	分散分析表								
9	自由度	変動	分散	観測された分散比	有意 F				
10	回帰	2	4.352675	2.176338	13.40544471	2.27185E-05			
11	残差	49	7.955017	0.162347					
12	合計	51	12.30769						
13	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95%	上限 95%	
14	切片	-0.99009	0.332037	-2.98186	0.00445355	-1.65734	-0.32283	-1.65734	-0.32283
15	都心	0.43141	0.113088	3.814488	0.000382858	0.204132	0.658668	0.204132	0.658668
16	年収(単位:10万円)	0.009613	0.002737	3.512126	0.000655585	0.004113	0.015113	0.004113	0.015113

P-値(危険率)がかなり小さいので回帰分析の実行はこれでストップする。  
説明変数と説明変数では説明変数選択基準の大きい説明変数を採用する。  
最適な判別予測式は次のようになる。

$$y = -0.90 + 0.431 \times \text{都心} + 0.010 \times \text{年収(単位:10万円)}$$

これに対して、No.53の各値(都心=1, 年収=10)に代入すると、DMレスポンスの有無の値は0.46となり、0.5より少ないのでレスポンスなし(=0)と予測される。実際の値は0なので、予想は的中しているといえる。

画面から戻る

図4 「DMのレスポンスの判別予測」の解析ストーリーをExcelで実行するための操作説明ページ

- 上田太一郎 (1998) データマイニング事例集. 共立出版.  
 上田太一郎 (1999) データマイニング実践集. 共立出版.  
 上田太一郎・菊田正雄・本田和恵 (2003) 実践ワークショップ Excel 徹底活用 多変量解析. 秀和システム.  
 淵上美喜・上田太一郎・古谷都紀子 (2004) 実践ワークショップ Excel 徹底活用 ビジネスデータ分析. 秀和システム.  
 森 裕一・山本義郎・宿久 洋・本多啓介 (2007) 教育・学習支援のためのデータ指向統計解析環境. 日本統計学会和文誌, 36(2), 327-347.  
 森 裕一・山本義郎・黒田正博・宿久 洋 (2008) 統計学習コンテンツの評価システム構築の試み. 2008 年度数学教育学会春季年会発表論文集, 146-148.  
 森 裕一・山本義郎・宿久 洋 (2007) 統計教育におけるデータ収集・登録機能の活用 - モバイル統計解析システム構築の試み -. 2007 年度数学教育学会春季年会発表論文集, 43-45.  
 森 裕一・山本義郎・宿久 洋 (2006) データ指向統計環境 DoSS@d における データ収集・登録機能. 第 34 回日本行動計量学会大会抄録集, 140-143.  
 宿久 洋・山本義郎・藤野友和・黒田正博・森 裕一 (2008) e ラーニングのための統計コンテンツ提供システム. 2008 年度統計関連学会連合大会報告集, 87.  
 山崎康博・森 裕一 (2005) データ指向型統計教育における統計学習システム. 2005 年度数学教育学会秋季例会発表論文集, 87-89.

付 録

表 4 「データマイニング事例集」の解析ストーリー

解析ストーリー	分野		専門分野				統計技法					手 法					
	大分類	小分類	マーケティング	生活と統計	経済学	生物統計	要約	視覚化	比較と検定	予測	分類	次元縮約	記述統計	回帰分析	PCA・CA	CLS・DISC	因子分析
白菜と白菜漬の関係	経	経		○	○			○		○				○	○		
ハードウェア仕様と最低価格の関係	経	経		○	○			○		○			○	○			
チラシ広告の効果の要因	経	経	○		○					○			○	○			
業種の判別	社	社						○						○			
OL 向け 1 泊観光旅行の評価	社	社								○						○	
カップヌードルの年代による好み	経	経	○		○			○	○					○			
パソコン販売の地域差	経	経	○		○			○	○								
年齢・性によるスポーツの好みの違い	入	入						○	○								
和菓子屋の比較	経	経	○		○			○	○						○		
消費者の値ごろ感	経	経	○		○			○	○					○			
支払っていいと思う外食代の評価	経	経	○	○			○						○				
缶コーヒーの好感度*	経	消	○		○			○					○				
月別売上高の予測	経	経	○		○			○		○				○	○		
パチンコ店数の予測	社	社						○		○				○	○		
経過年と販売量の関係	経	消	○		○			○		○				○	○		
CO <sub>2</sub> 量の予測	科	環						○		○				○	○		
スーパーと小売店の価格の違い	経	経	○	○	○									○	○		
調味料のスーパーと小売店の価格差	経	消	○	○	○			○						○	○		
プログラム出荷後のバグ数の予測	工	工						○					○				
パソコン価格の予測	経	消	○	○	○			○		○				○	○		
牛乳の売り上げ本数の予測	経	消	○	○	○			○		○				○	○		
パンの売上個数の予測	経	消	○	○	○			○		○				○	○		
パソコンの電気代の予測	経	消		○	○			○		○				○	○		
乗用車の保有台数の予測	経	経	○		○			○		○				○	○		
販促効果の評価	経	消	○		○			○		○				○	○		
好まれるスキー場の要因	経	消	○		○			○		○				○	○		
ガソリンスタンドの評価	経	消	○		○			○						○	○		
食器用コンパクト洗剤の評価	経	消	○	○	○			○						○	○		
栄養ドリンク剤の評価	経	消	○	○	○			○						○	○		
冷凍食品の評価	経	消	○	○	○			○						○	○		
女性が選ぶ海外ブランドの人気	経	消	○	○	○			○						○	○		
女性が選ぶ焼き立てパンの人気	経	消	○	○	○			○						○	○		
小売店への要望・不満の評価	経	消	○	○	○			○						○	○		
セルフ式コーヒー店の人気	経	消	○	○	○			○						○	○		
レジャー型プールの人気	経	消	○	○	○			○						○	○		
チュアブル錠胃腸薬の普及の要因	経	消	○	○	○			○						○	○		
MD デッキ搭載ミニコンポの評価	経	消	○	○	○			○						○	○		
DM のレスポンス率を高める要因	経	消	○		○			○		○				○	○		

PCA: 主成分分析, CA: コレスポネンダ分析, CLS: クラスタ分析, DISC: 判別分析  
 \* すでに、解析ストーリーとして DoSS@d に収録されているもの

表5 「データマイニング実践集」の解析ストーリー

解析ストーリー	分野		専門分野				統計技法					手法					
	大分類	小分類	マーケティング	生活と統計	経済学	生物統計	要約	視覚化	比較と検定	予測	分類	次元縮約	記述統計	回帰分析	PCA・CA	CLS・DISC	因子分析
会社の分類1	経	経			○			○								○	
会社の分類2	経	経			○			○								○	
会社の分類3	経	経			○							○					
晴の日と雨の日での製品の不良率	工	工											○	○			
触媒と製造温度の違いによる生成量	工	工								○			○	○			
最適条件時の特性値を予測	工	工											○	○			
調査法と謝礼金による回収率の違い	社	社											○	○			
製法と作業員による生産数の違い	工	工											○	○			
材料と熱処理条件による変形量	工	工								○			○	○			
ダイエット法の効果	医	健		○		○							○	○			
生成量の増減に影響を与える要因	工	工											○	○			
3元配置実験の分析	工	工											○	○			
上昇株の見分け方	経	経	○		○				○				○	○			
6漁港の水揚げ量の比較	経	経			○				○				○	○			
境港の水揚げの変化	経	経	○		○				○				○	○			
定番人気メニューの注文数の予測	経	消		○	○					○			○	○			
スーパーの客数予測*	経	経	○	○	○					○			○	○			
家庭のガス代の予測	経	消		○	○				○				○	○			
売上げ数の予測	経	経		○	○				○				○	○			
リフレッシュ休暇旅行のコンセプト	経	経	○		○					○			○	○			
情報リテラシー	経	経	○	○	○				○	○			○	○			
賃料の予測	経	消	○	○	○				○				○	○			
旅行業界の12か月先を予測	経	経	○	○	○				○				○	○			
旅行会社の旅行取扱高	経	経	○	○	○				○				○	○			
百貨店の売上高の抜高	経	経	○	○	○				○				○	○			
貯蓄理由の年代別の際	経	経	○	○	○				○				○	○			
女性とレジャー、外食への行動(女性)	経	消	○	○	○				○			○	○			○	○
家族と一緒にする外食	経	消	○	○	○				○				○	○			
株価の予測	経	経	○	○	○				○				○	○			
家計支出に占める交通・通信費の予測	経	消	○	○	○				○				○	○			
百貨店の売上の季節性の予測	経	経	○	○	○				○				○	○			
好まれるモバイルPCの要因	経	消	○	○	○				○				○	○			
好まれる定食の要因	経	消	○	○	○				○				○	○			
マンションの契約率の予測	経	消	○	○	○				○				○	○			
完全失業率の予測	社	社							○				○	○			
広告扱い高の予測	経	消							○				○	○			
新設住宅の着工予測	経	経	○		○				○				○	○			

PCA：主成分分析，CA：コレスポンデンス分析，CLS：クラスター分析，DISC：判別分析

\*すでに、解析ストーリーとしてDoSS®dに収録されているもの



表6 「実践ワークショップ Excel 徹底活用 多変量解析」の解析ストーリー

解析ストーリー	分野		専門分野				統計技法					手法					
	大分類	小分類	マーケティング	生活と統計	経済学	生物統計	要約	視覚化	比較と検定	予測	分類	次元縮約	記述統計	回帰分析	PCA・CA	CLS・DISC	因子分析
白菜と白菜漬の関係	経	経		○	○			○		○			○	○			
プロ野球選手の分類	社	他						○					○				
温泉旅館のサービスの分類	社	社	○	○				○				○					
企業別「技術力」「親近感」の比較	社	社		○	○			○				○					
キャンブル型レジャーの比較	社	社		○	○			○	○								
チェーン店評価データ	社	社		○				○									
データのグループ化	社	社		○				○									
価格とシェアの関係	経	経		○	○	○		○	○								
食料品会社の比較	経	経	○	○	○			○	○	○							
食器乾燥機の機能と初月販売数	経	経	○	○	○			○	○	○							
キャンブル型レジャーの比較	社	社		○				○	○					○			
選手の打率と出塁率	社	社		○				○	○								
従業員数・資本金・売上高の関係	経	経	○		○			○		○				○	○		
マンション価格を左右する要因	経	経			○			○					○	○			
出塁率を決める要因	経	経			○			○					○	○			
食器乾燥機の販売数に影響する要因	経	経	○		○			○		○			○	○			
アパート賃料の判別	経	経	○	○	○			○		○			○	○			
DMのレスポンスの判別予測	経	経	○		○			○		○			○	○			
金鉱か否かの判別	他	他						○		○			○	○			
売上を左右するアイテム	経	消	○		○			○		○			○	○			
売上に効いているアイテム	経	消	○		○			○		○			○	○			
宣伝効果に効いている要因	経	消	○		○			○		○			○	○			
DMのレスポンス率を高める要因	経	消	○		○			○		○			○	○			
社内旅行プランの評価	社	社						○		○			○	○			
好まれるスキー場の要因	経	経	○		○			○		○			○	○			
料理学校の人気の要因	経	経	○		○			○		○			○	○			
営業担当の性格の分類	社	社						○					○	○			
宿泊旅行の年代別の好み	社	社						○					○	○			
家でよく飲むお茶の種類	社	社						○					○	○			
キャンブルイメージの比較	社	社		○				○					○	○			

PCA：主成分分析，CA：コレスポネンダ分析，CLS：クラスター分析，DISC：判別分析

表7 「実践ワークショップ Excel 徹底活用 ビジネスデータ分析」の解析ストーリー

解析ストーリー	分野		専門分野				統計技法					手法					
	大分類	小分類	マーケティング	生活と統計	経済学	生物統計	要約	視覚化	比較と検定	予測	分類	次元縮約	記述統計	回帰分析	PCA・CA	CLS・DISC	因子分析
東京と大阪の百貨店の売り上げ	経	消	○	○	○			○					○				
定食チェーンのアンケート	経	経	○		○			○					○				
プレゼン資料の製作費のアンケート	社	社		○				○					○				
中学生の塾代に対する意識	社	社		○				○					○				
モーニングセットの値段への意識	社	社		○				○					○				
専有面積と賃料の関係	経	経		○	○			○		○			○	○			
経過年と賃料の関係	経	経		○	○			○					○	○			
ゲーム各社の連結業績	経	経			○			○					○	○			
売上高と経常利益の関係	経	経			○			○		○			○	○			
株価の予測	経	経			○			○		○			○	○			
新商品の売上高の予測	経	経		○	○			○		○			○	○			
社員数の予測	社	社						○		○			○	○			
曜日と需要高の関係	経	経	○		○			○					○	○			
経過月と月別客数の関係	経	経	○		○			○					○	○			
お中元の販促効果	経	経	○		○			○		○			○	○			
ペットに最善のリフォームプラン	経	経	○	○	○			○		○			○	○			
売れ筋商品と在庫管理*	経	消	○	○	○			○					○	○			
販売実績の評価	経	経	○		○			○		○			○	○			

PCA：主成分分析，CA：コレスポネンダ分析，CLS：クラスター分析，DISC：判別分析

\*すでに、解析ストーリーとしてDoSS@dに収録されているもの

## Statistical Education using Social Information Data

— To Enrich Contents in a Database of Example Analyses —

Yuichi MORI and Akihiro OHTA\*

*Department of Socio-Information,*

*Faculty of Informatics, Okayama University of Science*

*\*Graduate School of Informatics, Okayama University of Science*

*1-1 Ridai-cho, Okayama 700-0005, Japan*

(Received September 30, 2008; accepted November 7, 2008)

A web-based statistical system “Data-oriented Statistical System” or DoSS@d has been developed and located at <http://mo161.soci.ous.ac.jp/@d/>, which archives data sets and the corresponding analysis stories (reports of how to analyze the data in practice) and perform statistical analysis on these data sets on the web. DoSS@d consists of DoDStat@d (data and story database), DoAStat@d (online analysis system) and DoLStat@d (learning courses).

The analysis stories and learning statistics using them are very useful but the number of stories is not enough and the fields of stories stored in DoDStat@d are not well-balanced. So we have started to collect a number of examples, classify them under statistical points and transfer them to DoSS@d analysis stories. We performed this based on four books, which include lots of real examples mainly from economics, business and data mining. Totally 123 stories have been created, which can be used for statistical education using DoSS@d system. Since we installed many analysis stories into DoSS@d from several fields, it is expected that statistical education using those analysis stories make learners have the ability of correct statistical analysis - selecting a suitable method, operating statistical packages correctly and then interpreting the results properly.

**Keywords:** data oriented statistical system; e-learning; data set; analysis story, we-based system.