



日本銀行ワーキングペーパーシリーズ

## 「分からない」という回答から分かること？ —短観「物価見通し」における無回答バイアス—

宇野洋輔\*

yosuke.uno@boj.or.jp

安達孔\*\*

kou.adachi@boj.or.jp

No.18-J-5  
2018年6月

日本銀行  
〒103-8660 日本郵便（株）日本橋郵便局私書箱 30号

\* 調査統計局

\*\* 調査統計局（現・松江支店）

日本銀行ワーキングペーパーシリーズは、日本銀行員および外部研究者の研究成果をとりまとめたもので、内外の研究機関、研究者等の有識者から幅広くコメントを頂戴することを意図しています。ただし、論文の中で示された内容や意見は、日本銀行の公式見解を示すものではありません。

なお、ワーキングペーパーシリーズに対するご意見・ご質問や、掲載ファイルに関するお問い合わせは、執筆者までお寄せ下さい。

商用目的で転載・複製を行う場合は、予め日本銀行情報サービス局 (post.prd8@boj.or.jp) までご相談下さい。転載・複製を行う場合は、出所を明記して下さい。

# 「分からない」という回答から分かること？ —短観「物価見通し」における無回答バイアス—\*

宇野洋輔<sup>†‡</sup>

安達孔<sup>§</sup>

2018年6月7日

## 概要

本稿は、短観の「物価見通し」における「分からない」という回答の中から、真の意味では「分かる」と推察される回答を機械学習の手法を用いて識別する。そのうえで、真の意味では「分かる」と識別された企業について、何らかの理由で回答されなかった反実仮想的な「物価見通し」を推定する。そして、その推定結果にもとづいて、無回答バイアスの大きさを評価する。分析の結果として、次の四点を指摘することができる。第一に、真の意味では「分かる」が、何らかの理由で「分からない」と回答した企業は、きわめて少ない。第二に、真の意味では「分かる」と識別された企業は、真の意味でも「分からない」と識別された企業に比べて、相対的に企業規模が小さく非製造業に多い。第三に、真の意味では「分かる」と識別された企業の反実仮想的な「物価見通し」は、公表集計値と統計的に有意に異なる。第四に、以上を踏まえると、短観の「物価見通し」の無回答バイアスは、統計的には無視できる。

JELコード：C55、E31

キーワード：物価見通し、PU分類、無回答バイアス

---

\*本稿は、筆者ら2名と高田耕平(企業統計グループ・現釧路支店)の計3名による短観における非標本誤差の検証プロジェクトの成果の一部である。本稿の作成にあたっては、青木浩介氏(東京大学)、桑原茂裕氏(以下、日本銀行)、関根敏隆氏、鈴木純一氏、二宮拓人氏、一上響氏、稲村晃希氏、榎本英高氏、中山興氏、吉羽要直氏から有益なコメントを頂いた。記して感謝したい。もちろん、あり得べき誤りは筆者らに属する。また、本稿に示される内容や意見は、筆者ら個人に属するものであり、日本銀行の公式見解を示すものではない。

<sup>†</sup>連絡先：yousuke.uno@boj.or.jp

<sup>‡</sup>日本銀行調査統計局経済統計課企業統計グループ

<sup>§</sup>日本銀行調査統計局経済統計課企業統計グループ(現松江支店)

## 1 はじめに

一般に、アンケート調査における「分からない」という回答には、次のふたつのケースがある。ひとつは、「分かる」と「分からない」との間の線引きが明確な場合である。たとえば、「量子コンピュータの仕組みを知っていますか？」という質問に対する「分からない」という回答がこれにあたる。もうひとつのケースは、「分かる」と「分からない」との間の線引きがいくらか曖昧な場合である。たとえば、「2018年のワールドカップの優勝国はどこになると思いますか？」という質問に対する「分からない」という回答がこれにあたる。後者のケースでは、「分からない」という回答と「ブラジル」という回答を異なる回答とみなすかどうかは、議論の余地がある。すなわち、「分からない」と回答した人の中には、「不確実性が高く予想は困難だが、最も優勝する確率が高い国をあえて挙げれば、ブラジルだろう」あるいは「サッカーに興味はないが、どうしても国名を答えなければならないのであれば、ブラジルだろう」と考えている人が存在する可能性がある。その場合には、「分からない」という回答は、「ブラジル」と異なる回答であるとはみなせない。

本稿は、わが国の代表的なアンケート調査である「全国企業短期経済観測調査(以下、短観)」における「分からない」という回答を扱う。具体的には、短観において、2014年3月調査から調査項目に追加された、「物価見通し」の項目における「分からない」という回答について議論する。短観の「物価見通し」は、「自社の販売価格」と「物価全般」について、1年後、3年後、5年後の見通しを尋ねたものであるが、これには、「分からない」という回答がかなりの割合で存在する。たとえば、「物価全般」の5年後では、40%の企業が「分からない」と回答している。逆にいえば、60%の企業は、「物価全般」の5年後を「分かる」と回答している。本稿では、この「分かる」と「分からない」との間の線引きは、いくらか曖昧であると考え、そして、上記のワールドカップの例に即していえば、「分からない」という回答の中に潜在的には「ブラジル」とみなせる回答がある場合、その企業の「真の回答」は「ブラジル」とであると仮定する。ただし、これとは逆に、「ブラジル」という回答の中に潜在的には「分からない」とみなせる回答がある場合、その企業の「真の回答」を「分からない」とみなすこともできるのも否定できない。この点については、最後に議論する。

本稿の議論は、ふたつの仮定に立脚する。ひとつは、「分からない」という回答は、多義的であるという仮定である。すなわち、「分からない」という回答の中には、真の意味では「分かる」と推察される回答が存在すると仮定する。もうひとつは、「分かる」という回答は、多義的でないという仮定である。すなわち、「分かる」という回答の中には、真の意味では「分からない」とみなせる回答は存在しないと仮定する。これらの仮定のもと

で、「分からない」という回答の中から、真の意味では「分かる」と推察される回答を識別することを試みる。本稿は、この問題を機械学習の文脈で捉える。「分からない」という回答が多義的であることは、機械学習の文脈では、「分からない」という「ラベル」が不完全であることを意味する。このとき、「分かる」という「ラベル」が多義的でないのであれば、これは、Liu *et al.* [17, 18]、Elkan and Noto [10] らによって提示された PU 分類問題とみなすことができる。PU 分類問題は、明示的な Negative ラベルデータが存在しない状況のもとで、Positive ラベルデータと Unabeled データから、Positive と Negative を分類する問題である<sup>1</sup>。PU 分類問題を解くために本稿で採用する手法は、機械学習の分野でよく知られた、サポートベクトルマシンとロジスティック回帰モデルを組み合わせたアルゴリズムである。結果として、「分からない」という回答は、真の意味で「分からない」という場合と、真の意味では「分かる」が、何らかの理由で「分からない」と回答した場合とに分類される。さらに、真の意味では「分かる」と識別された回答については、何らかの理由で回答されなかった反実仮想的な「物価見通し」を明らかにする。その結果にもとづいて、「物価見通し」における無回答バイアスの大きさを評価する<sup>2</sup>。

本稿の貢献は、次の二点である。第一は、短観の「物価見通し」の無回答バイアスについて定量的に評価したことである。言うまでもなく、本稿が示す無回答バイアスの大きさは、特定の仮定のもとで計算したひとつの試算値であるため、それにもとづく評価も幅をもってみる必要がある。本稿の分析結果によれば、真の意味では「分かる」が、何らかの理由で「分からない」と回答した企業は、きわめて少ない。逆にいえば、「分からない」と回答したほとんどすべての企業は、真の意味において「分からない」と考えられる。それでも、非常に少ない割合ではあるが、真の意味では「分かる」と識別された企業が存在する。これらの企業は、真の意味でも「分からない」と識別された企業に比べて、相対的に企業規模が小さく非製造業に多い。本稿の推定結果によれば、これらの企業が有する反実仮想的な「物価見通し」は、公表集計値と統計的に有意に異なる。以上を踏まえると、短観の「物価見通し」の無回答バイアスは、統計的には無視できると考えられる。

第二の貢献は、無回答バイアスの問題を機械学習の文脈で捉え直したことである。無回答バイアスや非標本誤差といったトピックは、一部の統計学者や統計作成当局によって長く議論されてきた。また、2016 年以降のわが国の統計改革においても、統計精度の向上

---

<sup>1</sup>これに対して、明示的に Positive ラベルデータと Negative ラベルデータが与えられているような状況のもとで、Positive と Negative を分類する問題は、PN 分類問題と呼ばれている。なお、より広い概念では、PN 分類は教師付き学習 (supervised learning)、PU 分類は半教師付き学習 (semi-supervised learning/partially supervised learning) に分類される。

<sup>2</sup>無回答バイアスとは、調査非協力やある項目に対する無回答によって生じる推計値の偏りである。短観における「分からない」はひとつの回答であり無回答ではないが、集計に含まれないという点で、推計値に対して無回答と同様の影響を及ぼす。これを踏まえ、本稿では、「分からない」という回答から生じ得るバイアスおよび回答が得られなかったことから生じ得るバイアスをあわせて無回答バイアスと呼ぶ。

は重要なトピックのひとつであると認識されている(統計推進会議 [4])。もっとも、こうした議論において、機械学習の分野で得られた知見を応用しようとする取り組みは、これまでのところ、なお限定的である。よく知られているように、機械学習の分野では、近年、理論およびアルゴリズムの両面において、知見の蓄積が加速度的に進んでいる。こうした分野の知見を応用できるのであれば、無回答バイアスなど統計精度に関する議論にも一層の進展が期待できる。

本稿の構成は、以下のとおりである。2節では、本稿と関連する先行研究を整理し、本稿の立ち位置を明確にする。3節では、本稿で用いるデータの定義を確認する。4節では、「物価見通し」における「分からない」という回答に関する基本的な事実整理を行う。5節では、本稿が解こうとする問題をPU分類問題として定式化する。6節では、「分からない」という回答が実際に多義的であること、すなわち、真の状態と異なる回答が確かに存在することをイベント・スタディによって示す。8節では、本稿で採用する、真の状態の識別アルゴリズムについて議論する。8節では、真の状態の識別結果を示し、続く9節で無回答バイアスの試算を行う。最後の10節はまとめである。

## 2 関連研究

本稿は、機械学習の手法を用いて無回答バイアスの問題にアプローチする。したがって、本稿は、無回答バイアスやそれと表裏一体の関係にある欠測値補完に関する研究に加え、機械学習関連の研究とも手法面で関連する。本節では、無回答バイアスと欠測値補完に関する先行研究について整理し、手法面の関連研究は、7.1節において整理する。

無回答バイアスや欠測値補完に関する議論は、Little and Rubin [16]によって包括的に整理されて以降、統計作成当局においても重大な関心が寄せられてきた。宇都宮・園田 [1]と平川・鳩貝 [6]は、本稿と同じ短観のデータを用いて、売上高や設備投資額といった年度計画項目における欠測値補完方法について議論している。また、内閣府 [5]も、同府が作成する「機械受注統計調査報告」の受注額における無回答バイアスについて分析を行っている。さらに、統計作成当局から個票データの提供を受けた研究者による分析もみられる。星野 [8]と土屋 [3]は、統計数理研究所が作成する「日本人の国民性調査」の個票データを用いて、調査不能バイアスに関する高度な分析を行っている。

これら先行研究と本稿との間には、決定的な違いがある。すなわち、これまでの無回答バイアスや欠測値補完の議論は、補完すべき「真の値」が存在することを前提としている。これと対照的に、本稿では、補完すべき「真の値」が存在しない場合も許容している。1節で議論したように、真の意味で「分からない」という状況は、そもそも補完すべき「真

の値」が存在しないことを意味している。この点は、短観の「物価見通し」という、ある意味で回答することが難しい項目に固有の問題かもしれないが、これまでの無回答バイアスや欠測値補完に関する議論を一般化しているとみることも可能である。

### 3 データ

本稿で用いるデータは、日本銀行調査統計局が四半期ごとに作成・公表する短観の個票データである。本稿で扱うのは、「物価見通し」の項目である。

短観の「物価見通し」には、「物価全般」と「自社の販売価格」の二種類がある。「物価全般」について、調査票には、「物価全般(消費者物価指数をイメージしてください)の前年比に関して、1年後、3年後、5年後はそれぞれ何%になると考えますか。貴社のイメージに最も近いものを、以下の選択肢(1~10)の中から選んで太枠内にご記入下さい」とあり、(1)前年比+6%程度以上(+5.5%以上)、(2)前年比+5%程度(+4.5%~5.4%)、以下1%刻みで、(10)前年比-3%程度以下(-2.6%以下)という10個の選択肢から回答することを求めている。ただし、イメージをもっていない場合には、(11)先行きについては不確実性が大きいから、(12)変動したとしても経営にほとんど影響がないため意識していないから、(13)その他、三つの選択肢のいずれかを選択することを求めている。本稿で扱う「分からない」という回答は、(11)(12)(13)のいずれかを選んだものである<sup>3</sup>。(1)から(10)までの回答を「分かる」と定義する。

「自社の販売価格」については、調査票には「貴社の主要製商品の国内向け販売価格または主要サービスの国内向け提供価格に関して、現在の水準と比べた1年後、3年後、5年後の価格の見通しに最も近いものを、以下の選択肢(1~10)の中から選んで太枠内にご記入ください」とあり、選択肢は、(1)現在の水準と比べ+20%程度以上(+17.5%以上)、(2)現在の水準と比べ+15%程度(+12.5%~17.4%)、以下5%刻みで、(9)現在の水準と比べ-20%程度以下(-17.6%以下)、(10)分からない、の10個である。「物価全般」と異なり、「自社の販売価格」においては、「分からない」という選択肢はひとつだけである。よって、(10)を選んだ回答を「分からない」、(1)から(9)を選んだ回答を「分かる」と定義する。

短観の調査対象は、資本金が2,000万円以上の企業で、一定精度の母集団推計値が得られるように抽出される。一度標本に組み入れられた調査対象企業は、原則として継続して調査されるため、短観のデータはパネル構造を有している。短観における「物価見通し」の調査は、2014年3月調査から実施されているため、本稿のサンプルは、2014年3

<sup>3</sup>この選択肢(11)(12)(13)の取り扱いの妥当性については、補論A節を参照。

月調査から 2017 年 9 月調査までの 15 四半期分で、企業数は 11,471 社、サンプルサイズは 159,558、アンバランスなパネルデータである。

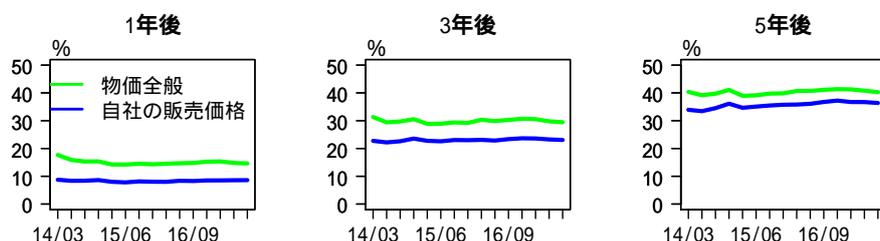
## 4 事実整理

本節では、「分からない」という回答に関する基本的な事実整理を行う。

### 4.1 時系列推移

図 1 は、「分からない」という回答割合の時系列推移を示したものである。図 1 から、次の三点を指摘することができる。第一に、「分からない」という回答割合は、時間を通じてほぼ一定である。第二に、「分からない」という回答割合は、いずれの年限でも、「物価全般」の方が「自社の販売価格」より高い。第三に、「分からない」という回答割合は、「物価全般」と「自社の販売価格」のいずれでも、年限が長くなるほど高くなる。

図 1: 「分からない」割合



### 4.2 属性別の「分からない」割合

表 1 は、企業規模と製造業・非製造業の別に、「分からない」と回答した企業の割合をみたものである。表 1 から、次の二点を指摘することができる。第一に、企業規模が大きくなるほど、「分からない」という回答割合が高くなる。たとえば、「物価全般」の 5 年後が「分からない」と回答した企業の割合は、資本金 1 億円以下の中小企業では 35%にとどまる一方、資本金 10 億円以上の大企業では 55%となっている。第二に、「分からない」という回答割合は、製造業の方が非製造業よりいくらか高い。

これらの傾向はいずれも、「物価全般」でも「自社の販売価格」でも、また、1 年後、3 年後、5 年後のいずれの年限でも観察することができる。これは、4.3 節でみるように、「分からない」という回答が、「物価全般」と「自社の販売価格」の間で、また、異なる年限の間で「入れ子構造」になっているためである。

表 1: 属性別の「分からない」割合

	物価全般			自社の販売価格		
	1年後	3年後	5年後	1年後	3年後	5年後
規模別						
大	0.25	0.44	0.55	0.16	0.37	0.49
中堅	0.14	0.29	0.40	0.07	0.22	0.35
中小	0.12	0.25	0.35	0.06	0.18	0.31
業種別						
製造	0.16	0.33	0.43	0.09	0.26	0.39
非製造	0.14	0.28	0.39	0.08	0.21	0.34

### 4.3 項目間および年限間の関係

ここでは、「物価全般」と「自社の販売価格」というふたつの項目間および1年後、3年後、5年後という三つの年限間での回答の関係を整理する。

項目間 表 2 は、「物価全般」と「自社の販売価格」というふたつの項目間の回答の分布を示したものである。

表 2: 「物価全般」と「自社の販売価格」の間の回答分布

	物価全般					
	分かる			分からない		
	1年後	3年後	5年後	1年後	3年後	5年後
自社の販売価格						
分かる	0.989	0.964	0.932	0.493	0.314	0.216
分からない	0.011	0.036	0.068	0.507	0.686	0.784
合計	1.000	1.000	1.000	1.000	1.000	1.000

表 2 の結果から、次の二点を指摘することができる。第一に、いずれの年限でも、「物価全般」が「分かる」場合には、「自社の販売価格」も「分かる」場合がほとんどである。すなわち、「物価全般」だけ「分かる」という回答はきわめて少ない。具体的には、「物価全般」だけ「分かる」という回答は、1年後では1.1%、5年後では6.8%となっている。第二に、「物価全般」が「分からない」場合には、「自社の販売価格」が「分かる」場合も「分からない」場合もある。ただし、年限が長くなるにつれて、「自社の販売価格」も「分からない」という回答の割合が上昇する。具体的には、「物価全般」が「分からない」という回答のうち、「自社の販売価格」も「分からない」という回答は、1年後では50.7%、5年後では78.4%となっている。

「物価全般」だけ「分かる」企業が非常に少ない一方、「自社の販売価格」だけ「分かる」企業は相応に存在する。それでは、「自社の販売価格」だけが「分かる」企業は、両方とも「分かる」企業と何が異なっているのだろうか。ひとつの試みとして、それぞれのグループの「自社の販売価格」見通しを比較してみる。

図 2: 「自社の販売価格」だけ「分かる」企業の「自社の販売価格」見通し

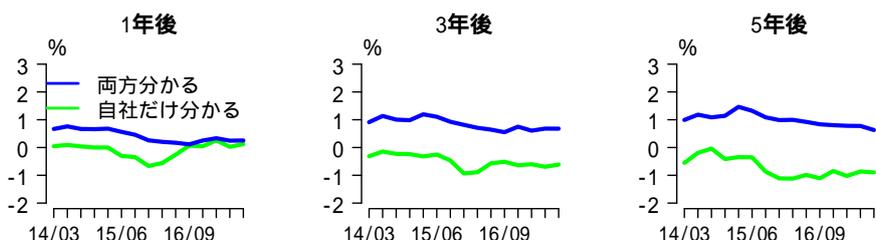


図 2 は、「自社の販売価格」だけが「分かる」企業 (図中の緑線) と、「自社の販売価格」も「物価全般」も両方「分かる」企業 (図中の青線) の「自社の販売価格」見通しを示したものである。ただし、企業規模の影響を取り除くために、資本金が 10 億円以上の大企業だけを集計している。興味深いことに、「自社の販売価格」だけが「分かる」企業の見通しは、両方とも「分かる」企業に比べて低い<sup>4</sup>。こうした傾向は、いずれの年限でも、確認することができる。また、年限が長くなるほど、見通しの差が大きくなっていく様子もみてとれる。ただし、「物価全般」が「分かる」かどうかによって、「自社の販売価格」見通しが異なるのか、あるいは逆に、「自社の販売価格」見通しの水準によって、「物価全般」が「分かる」かどうかの影響されるのか、それともただの見せかけの関係に過ぎないのか、この図から読み取ることはできない。

年限間 年限ごとの「分からない」という回答には、単調欠測 (monotone missing) の傾向がある。すなわち、1 年後が「分からない」場合には、3 年後および 5 年後も「分からない」場合が多く、3 年後だけ「分かる」場合や 5 年後だけ「分かる」という回答はきわめて少ない。具体的には、単調欠測パターンに当てはまらない回答の数は、「物価全般」で 43、「自社の販売価格」で 66、それぞれサンプル全体の 0.03% および 0.04% にとどまる<sup>5</sup>。

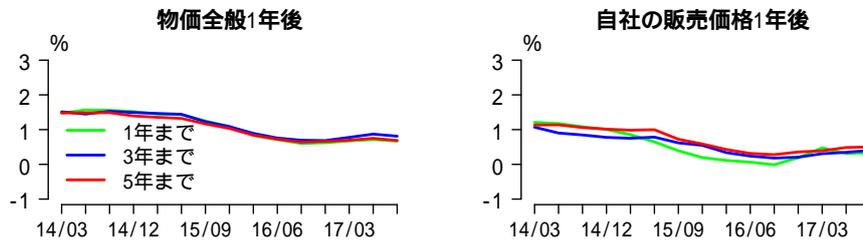
それでは、単調欠測パターンを満たす企業のうち、ある年限まで「分かる」企業は、それより先の年限まで「分かる」企業と何が異なっているのだろうか。ここでも、それぞれのグループの「物価見通し」を比較してみる。

図 3 は、1 年後だけ「分かる」企業 (図中の緑線)、1 年後と 3 年後が「分かる」企業 (図

<sup>4</sup>図 2 は大企業の集計結果だが、中堅企業と中小企業についても、この傾向は同様である。

<sup>5</sup>ここでの単調欠測は、(1) すべての年限で「分かる」、(2) 5 年後だけ「分からない」、(3) 3 年後と 5 年後が「分からない」、(4) すべての年限で「分からない」、という四つの場合と定義した。

図 3: 「分かる」年限が異なるグループごとの「物価見通し」



中の青線)、1年後と3年後と5年後が「分かる」企業(図中の赤線)、という三つのグループについて、各グループが共通して有する1年後の見通しを比較したものである。これを見ると、「物価全般」でも「自社の販売価格」でも、1年後の見通しは、どの年限まで「分かる」かによって、ほとんど異ならない。このことは、ある年限の「物価見通し」に関して、その年限まで「分かる」企業は、それより先の年限まで「分かる」企業と同じ情報を有している可能性が高いことを示唆している。

## 5 問題の定式化

本稿は、短観における「分からない」という回答の背後にある、「真の状態」を明らかにしようとする。これは、機械学習の分野で広く知られているPU分類問題として定式化することができる。PU分類問題は、Liu *et al.* [17, 18]、Elkan and Noto [10]らによって提示された二値分類問題のひとつで、明示的な Negative ラベルデータが存在しない状況のもとで、Positive ラベルデータと Unlabeled データから、Positive と Negative を分類する問題である。例を挙げれば、PU分類問題とは、A 大学に合格した受験生の属性データ (Positive ラベルデータ) と、A 大学を受験したが、合否が不明の受験生の属性データ (Unlabeled データ) から、ある受験生 B さんの A 大学の合否判定 (Positive と Negative の分類) を行う問題である。この問題が難しいのは、A 大学を受験して合格しなかった受験生の属性データ (Negative ラベルデータ) が利用可能でない状況のもとで、受験生 B さんの合否判定をしなければならない点にある。

今、真の状態において、物価見通しが「分かる」か否かを表す確率変数を  $y = \{1, 0\}$ 、短観において「分かる」という回答が観察されるか否かを表す確率変数を  $s = \{1, 0\}$  とする。ここで、 $y = 1$  が真の状態における「分かる」、 $s = 1$  が短観における「分かる」という回答を意味している。 $y$  は、観察することができない変数である。また、企業の属性を表す  $k$  個の確率変数からなるベクトルを  $x = (x_1, \dots, x_k)$  とする。これらの確率変数の間に、次のふたつの仮定を置く。ひとつは、短観における「分からない」という回答は、多

義的であるという仮定である。すなわち、「分からない」という回答には、真の状態における「分かる」が含まれていると考える。次のように書くことができる。

$$\Pr(s = 0 \mid \boldsymbol{x}, y = 1) > 0 \quad (1)$$

もうひとつは、「分かる」という回答は、多義的でないという仮定である。すなわち、「分かる」という回答には、真の状態における「分からない」は含まれていないと考える。次のように書くことができる。

$$\Pr(s = 1 \mid \boldsymbol{x}, y = 0) = 0 \quad (2)$$

1節でも述べたように、本稿では、「物価見通し」における「分かる」と「分からない」との間の線引きは、いくらか曖昧であると考えている。(1)式の仮定は、そうした本稿の考えをそのまま表現している。他方、(2)式はいささか強めの仮定だが、本稿では、回答する企業にとって、「分からない」にもかかわらず「分かる」とする誘因がないことから、この仮定を置いている<sup>6</sup>。

以上のふたつの仮定を置くと、真の状態における「分かる」 $y = 1$ の一部が短観において「分かる」という回答  $s = 1$  として観察され、真の状態における「分かる」 $y = 1$ の一部と「分からない」 $y = 0$ は、短観において「分からない」という回答  $s = 0$ として観察されることになる。別の言い方をすると、「分からない」という回答をデータの欠測と考えれば、本稿の問題設定は、短観の「物価見通し」において、「ランダムでない欠測 (missing not at random, MNAR)」が生じていると仮定していることになる<sup>7</sup>。なお、「分からない」という回答をデータの欠測とみなす場合には、この問題の構造は、Heckman [12]によって問題提起された、サンプル選択バイアスの問題と同じものである。

こうしたもとで、本稿が解こうとする問題は、「分かる」と回答した  $s = 1$ なる企業の属性情報を利用して、「分からない」と回答した  $s = 0$ なる企業  $i$ と  $j$ を、 $U_N = \{i \mid s_i = 0, y_i = 0\}$ と  $U_P = \{j \mid s_j = 0, y_j = 1\}$ という、ふたつの集合に分類するPU分類問題である。PU分類問題を解くことができれば、企業  $j \in U_P$ が有する反実仮想的な「物価見通し」を推定することができる。本稿では、その推定結果にもとづいて、無回答バイアスの大きさを評価する。

<sup>6</sup>本稿とは逆に、(1)式が等号で成り立ち、(2)式が不等号で成り立つと仮定すれば、NU分類問題となる。また、(1)式も(2)式も不等号で成り立つと仮定すれば、教師なし分類問題となる。

<sup>7</sup>欠測メカニズムの分類については、星野 [7] や内閣府 [5]などを参照。

## 6 真の状態と異なる回答の発見

本節では、真の状態の識別に先立って、5節の(1)式が短観の回答データによって支持されることをイベント・スタディによって確認する。

### 6.1 イベント・スタディの枠組み

真の世界では、先行きの物価を予想するコストの増加や企業内のリソースの減少、あるいは、先行きの物価についての不確実性が高まることなどによって、「分かる」から「分からない」への状態遷移が生じると考えられる。たとえば、資源価格の先行きが極端に不確実になる場合には、これまで「物価見通し」を「分かる」としてきた企業でも、「分からない」という状態に遷移する可能性がある。これは、真の状態における「分かる」から「分からない」への遷移である。

もっとも、そうした真の状態遷移は、短観の回答窓口が誰か、という事実とは、本来、無関係であると考えられる。すなわち、回答窓口が変更されたことによって、先行きの物価についての不確実性が高まるわけではないし、逆に、先行きの物価についての不確実性が高まったことによって、短観の回答窓口が変更になるといったことも同様に考えにくい。したがって、企業の「物価見通し」の真の状態は、回答窓口の変更と独立であると考えられる。これを5節での表記に即せば、真の状態遷移は、次のように書くことができる。

$$\Pr(y_t = 0 \mid \mathbf{x}_t, y_{t-1} = 1, z_t = 0) = \Pr(y_t = 0 \mid \mathbf{x}_t, y_{t-1} = 1, z_t = 1) \quad (3)$$

$t$  は、調査時点を表す。 $z = \{1, 0\}$  は短観の回答窓口が変更された場合に1、回答窓口が不変の場合にゼロをとる確率変数である。 $z$  は、真の状態  $y$  に影響を及ぼさないと考えているため、等号が成り立つ。

他方、もし、(1)式が成り立たないのあれば、すなわち、 $\Pr(s = 0 \mid \mathbf{x}, y = 1) = 0$  であれば、短観の回答  $s$  は真の状態  $y$  に一致する。これと(3)式から、次が成り立つ。

$$\Pr(s_t = 0 \mid \mathbf{x}_t, s_{t-1} = 1, z_t = 0) = \Pr(s_t = 0 \mid \mathbf{x}_t, s_{t-1} = 1, z_t = 1) \quad (4)$$

逆にいえば、(4)式が成り立たないのであれば、(1)式が成り立つことになる。このことは、短観の回答窓口の変更をひとつのイベントとして、「分かる」という回答から「分からない」という回答への遷移確率を比較すれば、(1)式が成り立つかどうかを確かめられることを示唆している。

## 6.2 イベント・スタディの結果

回答窓口の変更イベントを用いたイベント・スタディの結果を報告する。日本銀行調査統計局では、調査対象企業の回答窓口に関して、役職や氏名といった情報を管理している。ここでは、回答窓口の変更を、役職が不変で氏名が変更された場合と定義した。すなわち、回答窓口が経理課長 A 氏から経理課長 B 氏へ変更された場合のみを「窓口変更」とし、経理課長 A 氏から財務部長 C 氏への変更は「窓口変更」に含めていない。これは、6.1 節の議論に即して言えば、窓口変更イベントの有無を表す変数  $z$  を、真の状態  $y$  と独立であると考えられるようにしておくための処置である。

表 3: 「分かる」から「分からない」への状態遷移確率

	物価全般		
	1 年後	3 年後	5 年後
窓口変更時 ( $z = 1$ )	0.032	0.059	0.061
[3792]	(0.027,0.037)	(0.052,0.067)	(0.053,0.069)
窓口不変時 ( $z = 0$ )	0.017	0.032	0.037
[142709]	(0.016,0.018)	(0.031,0.032)	(0.036,0.038)
差	0.015	0.027	0.024
	自社の販売価格		
	1 年後	3 年後	5 年後
窓口変更時 ( $z = 1$ )	0.021	0.042	0.051
[3792]	(0.017,0.026)	(0.036,0.049)	(0.044,0.058)
窓口不変時 ( $z = 0$ )	0.013	0.027	0.037
[142709]	(0.011,0.016)	(0.026,0.028)	(0.036,0.038)
差	0.008	0.015	0.014

(注) 表中の () 内は、ブートストラップ法 (パーセントイル法、窓口変更時は反復 2000 回、窓口不変時は反復 200 回) によって計算した 95%信頼区間、[] 内は観測数を示す。

表 3 のイベント・スタディの結果から、次の二点を指摘することができる。第一に、「分かる」から「分からない」への状態遷移確率は、回答窓口の変更時において統計的に有意に高くなる。たとえば、「物価全般」の 1 年後をみると、ある時点において「分かる」と回答した企業が 3 か月後に「分からない」と回答する確率は、窓口が不変の場合には 1.7% である。この確率は、3 か月後に回答窓口が変更される場合には 3.2% と、統計的に有意に高くなる。この窓口変更時における統計的に有意な変化は、「自社の販売価格」でも「物価全般」でも、また、1 年後、3 年後、5 年後のいずれの年限でも確認することができる。これは、いずれの項目や年限でも、(1) 式が成り立っていることを示唆している。

第二に、状態遷移確率の差は、年限が長いほど大きく、また、「物価全般」の方が「自社の販売価格」より大きい。たとえば、「物価全般」における状態遷移確率の差をみると、1年後では1.5%ポイントである一方、5年後では2.4%ポイントとなっているほか、5年後における状態遷移確率の差をみると、「自社の販売価格」では1.4%ポイントである一方、「物価全般」では2.4%ポイントとなっている。

表 3 の結果は、窓口変更の有無  $z$  が短観の回答  $s$  と相関していることを示唆している。すなわち、1 節の例に即していえば、真の状態  $y$  が同じであったとしても、ある担当者は「ブラジル」と回答し、後任の別の担当者は「分からない」と回答する場合が確かにあることを示唆している。こうした回答の「揺れ」は、回答窓口の役職とは関係のない担当者に固有の要因によるものである。たとえば、慎重さやアンケート調査に対するスタンスなどが考えられる。ただし、繰り返しになるが、「ブラジル」という回答と「分からない」という回答のどちらを「真の回答」とするかは、どのような仮定を置くかによる。したがって、このイベント・スタディの結果をもって、窓口変更前後のどちらの回答が「正しい」のかを議論することは適当ではない。

なお、このイベント・スタディの結果は、5 節の (1) 式が成り立つことを支持する一方、もうひとつの仮定である (2) 式については、何も示唆しないことに留意されたい。

## 7 真の状態の識別アルゴリズム

6 節では、イベント・スタディによって、「分からない」という回答が実際に多義的であることを定量的に確認した。本節では、「分からない」という回答  $s = 0$  の背後にある真の状態  $y = \{0, 1\}$  を識別するアルゴリズムについて議論する。以下では、 $U = \{U_N, U_P\}$  とし、「分かる」と回答した企業  $h$  の集合を  $P = \{h \mid s_h = 1, y_h = 1\}$  と表記する。

### 7.1 PU 分類問題を解くアルゴリズム

Yang *et al.* [22] の整理にもとづくと、これまでに提案されてきた PU 分類問題を解くアルゴリズムは、大別して、次の四種類である。第一は、サンプル選択メカニズム  $\Pr(s = 1 \mid y = 1)$  に何らかの仮定を置いて、 $\Pr(y = 1 \mid \boldsymbol{x})$  を直接推定するものである (Heckman [12]、Lee and Liu [14]、Elkan and Noto [10] など)。本稿の場合、サンプル選択メカニズムに関して、先験的に明らかでないことが存在しない。したがって、 $\Pr(s = 1 \mid y = 1)$  に特定の仮定を置くことができない。

第二は、ブートストラップ法にもとづくアルゴリズムである (Mordelet and Vert [19]、Yang *et al.* [22] など)。これは、集合  $U$  からランダムにサンプリングしたデータを Negative ラベルデータとみなして PN 分類問題を解くことを繰り返すもので、Breiman [9] のアイデアを PU 分類問題に援用したものである。このアルゴリズムでは、集合  $U$  内の潜在的な Positive ラベルデータと集合  $P$  内の Positive ラベルデータが同じ確率分布に従うことを仮定する。これは、本稿の文脈では、無回答バイアスがないことを仮定することと同義である。したがって、このアルゴリズムをそのまま用いることはできない。

第三は、「2ステップ戦略」と呼ばれる、ヒューリスティックなアルゴリズムである (Liu *et al.* [17, 18] など)。このアルゴリズムでは、ステップ1において、集合  $U$  内のデータから高い確度で Negative であると見込まれるデータ (reliable negative,  $RN \subset U_N$ ) を見つけ出し、集合  $P$  内の Positive ラベルデータと集合  $RN$  内のデータを用いて PN 分類問題を解く。すなわち、ステップ1において、PU 分類問題を PN 分類問題に変換する。続くステップ2では、ステップ1の PN 分類問題を解いて得られた分類器を用いて、集合  $U$  内のデータを  $U_N$  と  $U_P$  に分類する。本稿の場合、「分からない」と回答した企業のうち、どの企業が真の意味で「分からない」と高い確度で見込まれる企業  $i' \in RN$  なのかは先験的に明らかではない。それでも、真の意味で「分からない」と高い確度で見込まれる企業さえ特定できれば、「2ステップ戦略」自体は、サンプル選択メカニズムに特定の仮定を置かないヒューリスティックなアルゴリズムであるため、本稿の PU 分類問題にも応用可能である。

第四は、「1クラス分類」と呼ばれる分野との問題の類似性を利用したものである (Li *et al.* [15] など)。すなわち、Positive ラベルデータだけを「教師なし訓練データ」として利用し、そこで得られた Positive ラベルデータが有する「境界線」を集合  $U$  内のデータに適用して、「境界線」の内側にあるデータを  $U_P$  に、外側にあるデータを  $U_N$  に分類するというアイデアである。Khan and Madden [13] によれば、「1クラス分類」で最もよく用いられるアルゴリズムは、Tax and Duin [20, 21] によって提示された、1クラス SVM (One-class Support Vector Machine) である。1クラス SVM は、集合  $U$  内の潜在的な Positive ラベルデータと集合  $P$  内の Positive ラベルデータが従う確率分布の形状を先験的に仮定する必要がない。また、通常の PU 分類問題では、僅かな Positive ラベルデータと大量の Unlabeled データを想定することが多く、その場合には、1クラス SVM は必ずしも適切でないと思われるが、本稿のように、Positive ラベルデータの数が相対的に多い場合には、同手法を用いることで望ましい結果が得られることが期待できる。

## 7.2 本稿のアルゴリズム

7.1 節の整理を踏まえ、本稿では、「2ステップ戦略」と1クラスSVMを組み合わせてPU分類問題にアプローチする。

**ステップ1** ステップ1では、1クラスSVMを用いて、真の意味で「分からない」と高い確度で見込まれる企業  $i' \in RN$  を特定する。7.1 節でみたように、本稿の場合、企業  $i' \in RN$  を特定するための先験的な知識があるわけではない。ここでは、1クラスSVMにより「クラス外」に分類された企業を、事後的な意味で、真の意味で「分からない」可能性が高いものとみなす。

まず、集合  $P$  内にある  $M$  先の企業の属性  $\mathbf{x}_h (h = 1, \dots, M)$  をカバーする超球 (hyper-sphere) を求める。具体的には、以下の最適化問題を解くことによって超球の中心  $a$  と半径  $R$  を求める。

$$\min \left[ R^2 + C \sum_h^M \xi_h \right], \text{ subject to } \|\mathbf{x}_h - a\|^2 \leq R^2 + \xi_h, \xi_h \geq 0, \forall h$$

ここで、 $\xi_h$  はスラック変数、 $C$  は正則化パラメータ、 $\|\cdot\|$  は  $l_2$  ノルムを意味している。

最適化問題を解いて得られた中心  $\hat{a}$  と半径  $\hat{R}$  の超球を用いて、集合  $U$  内の企業を分類する。直観的にいえば、超球の表面および外側に分類された集合  $U$  内の企業は、集合  $P$  内の「分かる」と回答した、いずれの企業とも明確に異なる属性を有しているといえる。本稿では、これを真の意味で「分からない」と高い確度で見込まれる企業  $i' \in RN$  とみなす。なお、 $l_2$  ノルムの計算には、Tax and Duin [21] によって高い分類精度を得られることが確認されている、ガウスカーネルによるカーネルトリックを採用する。

**ステップ2** ステップ2では、ステップ1で得られた集合  $RN$  を繰り返し計算によって少しずつ拡張しながら、集合  $U \setminus RN$  内のデータを  $U_P$  と  $U_N$  に分類していく。具体的には、集合  $P$  内の  $L$  先の「物価見通し」を Positive ラベルデータ、ステップ1で集合  $RN \subset U_N$  に分類された企業  $i' \in RN$  の「物価見通し」を Negative ラベルデータとして、PN分類問題を解く。本稿では、以下のように、 $\Pr(y = 1 | \mathbf{x})$  をロジスティック回帰モデルによって特定化する。

$$\Pr(y = 1 | \mathbf{x}) = [1 + \exp(-\{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k\})]^{-1}$$

推定手法は最尤法である。推定の結果として得られた分類器  $\hat{\Pr}(y = 1 | \mathbf{x})$  を集合  $U \setminus RN$  内の企業に適用し、 $\hat{\Pr}(y = 1 | \mathbf{x}) < 0.5$  を満たす企業を集合  $RN$  に加える。そ

して、Positive ラベルデータのサイズ  $L$  を固定したうえで、新たに拡張された集合  $RN'$  内の Negative ラベルデータとの間で PN 分類問題をもう一度解く。このプロセスを集合  $U \setminus RN$  内に  $\Pr(y = 1 | \boldsymbol{x}) < 0.5$  を満たす企業が存在しなくなるまで繰り返す<sup>8</sup>。

### 7.3 ハイパーパラメータの設定

機械学習の分野では、学習によって推定値を得られないパラメータをハイパーパラメータと呼んでいる。本稿のアルゴリズムでも、いくつかのハイパーパラメータを設定しなければならない。

**ステップ 1** まず、1 クラス SVM の正則化パラメータ  $C$  を設定する。よく知られているように、1 クラス SVM の分類結果は、 $C$  の大きさによって大きく異なってしまふ。たとえば、 $C$  を小さく設定すれば、サポートベクトルが増加するため、本稿の場合には、集合  $RN$  に分類される企業が多くなると見込まれる。もっとも、本稿のアルゴリズムでは、ステップ 2 において、 $RN$  を少しずつ拡張していくため、ステップ 1 での  $C$  の設定によって分類結果が決まるわけではない。本稿の目的に照らせば、 $RN$  をできる限り正しく識別できるように、 $C$  を大きめに設定することが望ましい。この点、Tax and Duin [21] が示すように、外れ値の割合を  $\nu$  とすると、 $C$  と  $\nu$  の間には、 $C = 1/\nu M$  という関係がある。これを踏まえ、本稿では、直観的な解釈が容易な  $\nu$  を小さめに設定する。具体的には、 $\nu = 0.02$  とした。このことは、集合  $P$  内に 2% の割合で存在する、他のいずれの企業とも属性の似ていない特異な企業と同じような特徴をもつ集合  $U$  内の企業を集合  $RN$  に分類することを意味している。 $RN$  をできる限り正しく識別できるようにという同じ理由から、ステップ 1 における、ふたつめのハイパーパラメータであるガウスクアーネルのバンド幅も小さめの値、具体的には 0.05 と設定した。

ステップ 1 では、可能な限り多くのデータを推定に用いることが望ましいと考えられるため、「分かる」と回答した、すべてのデータを用いて 1 クラス SVM の推定を行う。すなわち、ステップ 1 における最後のハイパーパラメータである  $M$  については、 $M = \#P$  とした<sup>9</sup>。

**ステップ 2** 次に、ステップ 2 で利用する Positive ラベルデータのサイズ  $L$  を設定する。ステップ 2 では、ステップ 1 で得られた集合  $RN$  に分類された企業を Negative ラベルデータ、集合  $P$  内のデータを Positive ラベルデータとして、ロジスティック回帰モデルの推定

<sup>8</sup>推定にはオープンソースのソフトウェアである **R** を利用した。コードは補論 **B** 節を参照。

<sup>9</sup> $\#P$  は、集合  $P$  に含まれる要素の数を意味する。

を繰り返す。もっとも、集合  $P$  内のすべてのデータを用いると、Negative ラベルデータに対して、Positive ラベルデータが多過ぎてしまう。この点、Liu *et al* [17] も議論しているように、ステップ 2 における Positive ラベルデータの最適なサイズは先験的に明らかではない。本稿では、ステップ 1 において集合  $RN$  に分類された企業数と同数を集合  $P$  内からランダムにサンプリングすることとした。すなわち、 $L = \#RN$  とした。2 回目以降のロジスティック回帰モデルの推定においては、Positive ラベルデータは固定されるため、繰り返し計算のたびに推定に用いるデータ内の Negative ラベルデータの割合が高くなっていくことになる。

## 7.4 説明変数

企業の属性を表す  $k$  個の説明変数は、業種 (短観の業種分類、カテゴリー数 : 31)、企業規模 (短観の母集団推計に用いられる企業規模分類、カテゴリー数 : 20)、所在地 (都道府県別、カテゴリー数 : 47)、DI の「最近」(13 項目×選択肢の数、カテゴリー数 :  $13 \times 4$ )、DI の「先行き」(8 項目×選択肢の数、カテゴリー数 :  $8 \times 4$ )、時間ダミー (15 四半期分、カテゴリー数 : 15) で、すべてカテゴリー変数である。これらをダミー変数として数えれば、説明変数の数は 192 となる。なお、この説明変数のセットは、ふたつのステップで同一である。

## 8 識別結果

本節では、7.2 節で示したアルゴリズムによる真の状態の識別結果を報告する。

### 8.1 識別結果

真の状態の識別結果は、表 4 のとおりである。

識別の結果は、ハイパーパラメータの設定に依存する面があるため、相応の幅をもってみる必要がある。それでも、ある程度には頑健な結果として、「分からない」と回答した企業のほとんどは、真の意味で「分からない」とみられることを指摘できる。ただし逆にいえば、非常に少ない数ながら、真の意味で「分かる」と識別された企業も存在している。これは、6.2 節のイベント・スタディの結果と整合的である。

表 4 の結果によれば、真の意味で「分からない」と識別された企業は、いずれの項目や年限でみても、98%ないし 99%となっている。これを調査回ごとにみれば、真に「分か

表 4: 識別結果

	物価全般					
	1 年後		3 年後		5 年後	
分かる (# $P$ )	135600	(0.85)	111884	(0.70)	95291	(0.60)
分からない (# $U$ )	23958	(0.15)	47674	(0.30)	64267	(0.40)
真に「分かる」 (# $U_P$ )	315	(0.01)	407	(0.01)	411	(0.01)
真に「分からない」 (# $U_N$ )	23643	(0.99)	47267	(0.99)	63856	(0.99)
	自社の販売価格					
	1 年後		3 年後		5 年後	
分かる (# $P$ )	146272	(0.92)	122817	(0.77)	102679	(0.64)
分からない (# $U$ )	13286	(0.08)	36741	(0.23)	56879	(0.36)
真に「分かる」 (# $U_P$ )	289	(0.02)	238	(0.01)	188	(0.00)
真に「分からない」 (# $U_N$ )	12997	(0.98)	36503	(0.99)	56691	(1.00)

(注) 表中の () 内の計数は割合。 $U_P$  と  $U_N$  の割合は  $U$  に対する割合。

る」と識別された企業はきわめて少なくなる。たとえば、「物価全般」の5年後でみると、真に「分かる」と識別された企業は、2014年12月調査ではわずかに16先である。

## 8.2 識別結果の妥当性

ここでは、8.1節で得た識別結果の妥当性について、業種と企業規模の面から確認する。

**業種** 表5は、識別されたグループごとに製造業と非製造業の割合を示したものである。

表5の結果から、次の二点を指摘することができる。第一に、いずれの項目や年限でみても、真に「分かる」グループ( $U_P$ )は、「分からない」グループ( $U$ )に比べて、非製造業の割合が高い。たとえば、「物価全般」の1年後でみると、「分からない」グループの非製造業の割合が56%である一方、真に「分かる」グループの同割合は75%となっている。このことは、製造業か否かによって真に「分かる」か否かがいくらか説明できることを意味している。

第二に、真に「分かる」グループの製造業と非製造業の割合は、「分からない」グループより「分かる」グループに近い。実際、「物価全般」の5年後でみると、真に「分かる」グループと「分かる」グループの非製造業の割合は61%で一致している。このことは、本稿のアルゴリズムが、「分からない」と回答した企業の中から、「分かる」企業と平均的にはよく似た業種属性のグループを真の意味で「分かる」と識別していることを意味している。その意味で、識別の結果は、妥当なものである可能性が高い。

表 5: グループごとの製造業と非製造業の割合

	物価全般					
	1年後		3年後		5年後	
	製造	非製造	製造	非製造	製造	非製造
「分かる」( $P$ )	0.40	0.60	0.39	0.61	0.39	0.61
「分からない」( $U$ )	0.44	0.56	0.44	0.56	0.43	0.57
真に「分かる」( $U_P$ )	0.25	0.75	0.35	0.65	0.39	0.61
真に「分からない」( $U_N$ )	0.45	0.55	0.44	0.56	0.43	0.57

	自社の販売価格					
	1年後		3年後		5年後	
	製造	非製造	製造	非製造	製造	非製造
「分かる」( $P$ )	0.40	0.60	0.39	0.61	0.39	0.61
「分からない」( $U$ )	0.45	0.55	0.46	0.54	0.44	0.56
真に「分かる」( $U_P$ )	0.42	0.58	0.43	0.57	0.29	0.71
真に「分からない」( $U_N$ )	0.45	0.55	0.46	0.54	0.44	0.56

企業規模 次に、それぞれのグループについて、企業規模を比較する。表 6 は、それぞれのグループの資本金の平均値を計算したものである。

表 6 の結果から、次の二点を指摘することができる。第一に、真に「分かる」企業の資本金の平均値は、おおむねすべての項目において、真に「分からない」企業に比べて統計的に有意に低い<sup>10</sup>。この結果は、企業規模の違いによって真に「分かる」か否かがいくらか説明できることを意味している。宇野・永沼・原 [2] が指摘したように、企業のインフレ予想には企業規模間での異質性が強い。ここでの結果は、その指摘と整合的である。ただし、「物価全般」の3年後と5年後については、真に「分かる」企業の資本金の平均値は、真に「分からない」企業より低いものの、その差は統計的に有意ではない。

第二に、真に「分かる」企業の資本金の平均値は、おおむねすべての項目において、「分かる」企業の資本金の平均値と統計的に有意に異なる。これは、本稿のアルゴリズムが、「分かる」企業と平均的には同水準の資本金規模の企業を真に「分かる」と識別していることを意味している。その意味で、識別の結果は、妥当であると考えられる。ただし、「自社の販売価格」の3年後と5年後については、真に「分かる」企業の資本金の平均値は、「分かる」企業に比べても、統計的に有意に低い。

<sup>10</sup> こうした傾向は、中央値でもみても同様である。

表 6: グループごとの資本金の平均値

	物価全般		
	1年後	3年後	5年後
「分かる」( $P$ )	2610 (2517,2703)	2088 (2012,2164)	1932 (1854,2010)
「分からない」( $U$ )	10805 (10082,11528)	7953 (7540,8366)	6670 (6356,6983)
真に「分かる」( $U_P$ )	2106 (957,3253)	3159 (48,6270)	4872 (533,9211)
真に「分からない」( $U_N$ )	10921 (10188,11654)	7994 (7578,8410)	6681 (6367,6996)
	自社の販売価格		
	1年後	3年後	5年後
「分かる」( $P$ )	2663 (2572,2755)	2174 (2098,2250)	2151 (2068,2234)
「分からない」( $U$ )	16798 (15544,18052)	9412 (8887,9937)	6890 (6544,7236)
真に「分かる」( $U_P$ )	5693 (1207,10178)	519 (201,837)	483 (182,785)
真に「分からない」( $U_N$ )	17045 (15767,18323)	9470 (8942,9998)	6911 (6564,7259)

(注) 表中の () 内は、正規分布を仮定した場合の 95%信頼区間。百万円。

## 9 無回答バイアス

本節では、8.1 節において真の意味で「分かる」と識別された企業について、それらの企業が有する反実仮想的な「物価見通し」を推定する。その推定結果にもとづいて、無回答バイアスの大きさを評価する。

### 9.1 反実仮想的な「物価見通し」

真の意味で「分かる」と識別された企業  $j \in U_P$  が有する反実仮想的な「物価見通し」は、企業  $j \in U_P$  と「分かる」と回答した企業  $h \in P$  をマッチングすることによって推定する。具体的には、7.4 節で示した説明変数のセットを再度利用し、ロジスティック回帰モデルにもとづく傾向スコア・マッチング (propensity score matching) を行う<sup>11</sup>。これによ

<sup>11</sup>傾向スコア・マッチングの詳細は、星野 [7] や内閣府 [5]などを参照。なお、マッチングには、R の MatchIt パッケージを利用した。

り、「分かる」と回答した企業の中から、真の意味で「分かる」と識別された企業  $j \in U_P$  に「最も近い」とみなすことができる企業  $h \in P$  を探索することができる。そして、マッチングの結果としてペアとなった企業  $h \in P$  の「物価見通し」を、企業  $j \in U_P$  の反実仮想的な「物価見通し」の推定値とする。

図4は、集合  $U_P$  内の企業が有する反実仮想的な「物価見通し」の平均値 (図中の赤線) と、正規分布を仮定した場合の95%信頼区間 (図中の点線) を示している。

図4: 反実仮想的な「物価見通し」

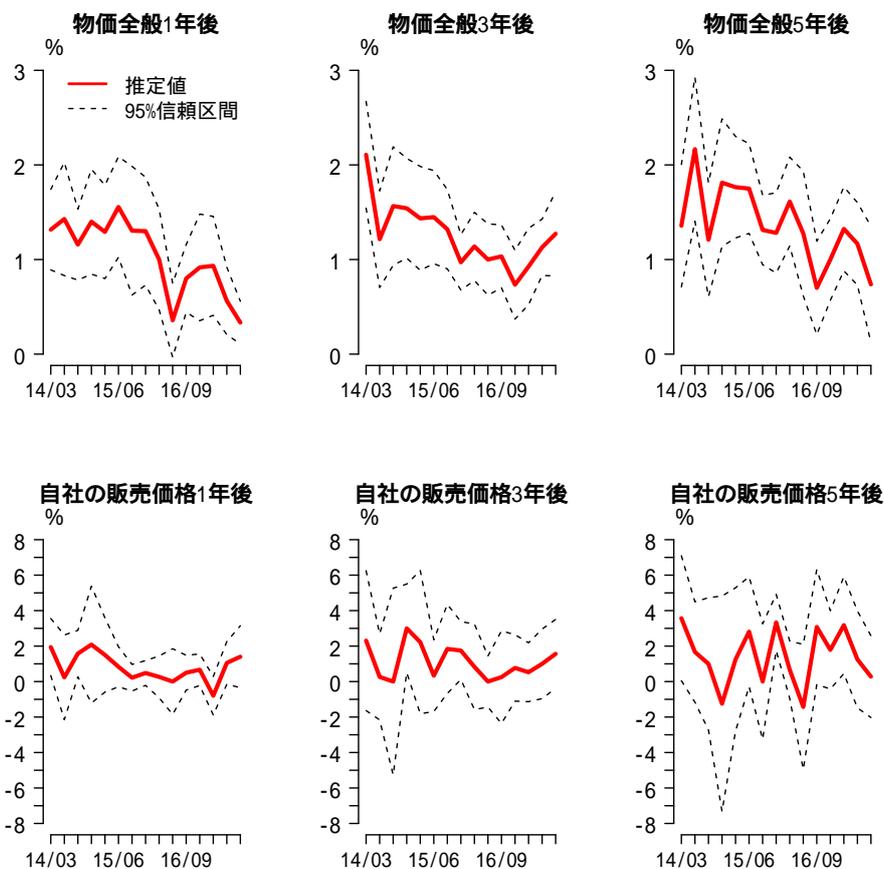


図4から、反実仮想的な「物価見通し」は、変動の幅が大きい点を指摘することができる。これは、真の意味で「分かる」と識別された企業が非常に少ないためである。また、「物価全般」でも「自社の販売価格」でも、年限が長くなるほど、変動の幅が大きくなる。宇野・永沼・原 [2] が示しているように、短観の「物価見通し」は、年限が長いほど標準偏差が大きい。したがって、反実仮想的な「物価見通し」においても、年限が長いほど標準偏差が大きくなるということは、真の意味で「分かる」と識別された企業の属性分布が「分かる」企業の属性分布と似ている可能性が高いことを示唆している。この点は、9.2節

で改めて議論する。

## 9.2 無回答バイアス

真の意味で「分かる」企業の反実仮想的な「物価見通し」の平均値を  $\mu_{UP}$ 、「分かる」企業の「物価見通し」の平均値(公表集計値)を  $\mu_P$ 、真の意味で「分かる」企業の割合を  $q$  と書くと、無回答バイアス  $B$  は、 $B = q(\mu_P - \mu_{UP})$  と定義される。ただし、 $q$  は、「分かる」企業と真の意味で「分かる」企業の合計に対する割合である。表 7 は、無回答バイアスの試算結果である。

表 7 の結果をみると、短観の「物価見通し」における無回答バイアスは、いずれの項目や年限でみても、ゼロとなっている。これには、ふたつの要因が寄与している。ひとつは、表 4 でみたように、真の意味で「分かる」企業が非常に少ない、すなわち、 $q$  が非常に小さいことである。もうひとつは、真の意味で「分かる」企業の反実仮想的な「物価見通し」が公表集計値と統計的に有意に異ならない、すなわち、 $(\mu_P - \mu_{UP})$  がゼロとみなせることである。 $(\mu_P - \mu_{UP})$  がゼロとみなせるということは、真の意味で「分かる」企業と「分かる」企業の属性分布が似ていることを示唆している。実際、8.2 節でみたように、製造業と非製造業の割合、資本金の平均値といった側面では、それらのグループはよく似ている。このことは、ある特定の属性を有する企業が「分かる」にもかかわらず、「分からない」と回答しているわけではないことを意味している。言い換えると、真の意味で「分かる」企業は、「分かる」企業と同じ確率分布からランダムに生じたとみなせる可能性が高い。1 節の例に即していえば、ランダムに選ばれたきわめて少数の企業が「ブラジル」と答えずに「分からない」と回答している状況であると推察される。

## 9.3 欠測値補完に関する議論

無回答バイアスと欠測値補完は、表裏一体の関係にある。ここでは、真に「分かる」と識別された企業に対する欠測値補完について議論する。

星野 [7, 8] や内閣府 [5] も強調するように、欠測値補完について意味のある議論をするためには、欠測メカニズムに何らかの仮定を置かなければならない。極端な例を挙げれば、欠測が「完全にランダム (missing completely at random, MCAR)」に生じるのであれば、欠測値補完そのものが必要ない。他方、「ランダムでない欠測」を仮定するのであれば、欠測したデータを何らかの方法で推定することが求められる。いずれにしても、欠測メカニズムに関する検討をないがしろにすれば、意味のある欠測値補完の議論はできない。

表 7: 無回答バイアス

物価全般									
	1年後			3年後			5年後		
	$\mu_P$	$\mu_{U_P}$	$B$	$\mu_P$	$\mu_{U_P}$	$B$	$\mu_P$	$\mu_{U_P}$	$B$
全体	1.1 (0.00)	1.1 (0.07)	0.0	1.3 (0.00)	1.2 (0.06)	0.0	1.4 (0.00)	1.4 (0.07)	0.0
規模別									
大	0.8 (0.01)	0.7 (0.13)	0.0	1.0 (0.01)	1.0 (0.11)	0.0	1.0 (0.01)	1.1 (0.13)	0.0
中堅	1.0 (0.01)	1.1 (0.14)	0.0	1.2 (0.01)	1.3 (0.11)	0.0	1.3 (0.01)	1.1 (0.15)	0.0
中小	1.2 (0.00)	1.2 (0.10)	0.0	1.5 (0.01)	1.3 (0.08)	0.0	1.5 (0.01)	1.6 (0.09)	0.0
業種別									
製造・素材	1.1 (0.01)	1.3 (0.21)	0.0	1.3 (0.01)	1.3 (0.17)	0.0	1.3 (0.01)	1.3 (0.16)	0.0
製造・加工	1.0 (0.01)	1.1 (0.16)	0.0	1.3 (0.01)	1.2 (0.11)	0.0	1.3 (0.01)	1.3 (0.15)	0.0
非製造	1.1 (0.00)	1.1 (0.08)	0.0	1.3 (0.00)	1.2 (0.08)	0.0	1.4 (0.01)	1.4 (0.09)	0.0
自社の販売価格									
	1年後			3年後			5年後		
	$\mu_P$	$\mu_{U_P}$	$B$	$\mu_P$	$\mu_{U_P}$	$B$	$\mu_P$	$\mu_{U_P}$	$B$
全体	0.6 (0.01)	0.7 (0.19)	0.0	1.3 (0.01)	1.1 (0.32)	0.0	1.6 (0.02)	1.5 (0.40)	0.0
規模別									
大	0.4 (0.02)	0.5 (0.60)	0.0	0.6 (0.03)	0.1 (0.84)	0.0	0.7 (0.04)	1.2 (0.66)	0.0
中堅	0.5 (0.01)	0.3 (0.26)	0.0	1.0 (0.02)	0.4 (0.50)	0.0	1.1 (0.03)	1.7 (0.79)	0.0
中小	0.8 (0.01)	1.0 (0.24)	0.0	1.7 (0.02)	1.8 (0.44)	0.0	2.2 (0.02)	1.5 (0.59)	0.0
業種別									
製造・素材	0.8 (0.02)	0.7 (0.53)	0.0	1.5 (0.04)	2.0 (0.90)	0.0	1.8 (0.05)	1.5 (0.66)	0.0
製造・加工	0.1 (0.02)	-0.3 (0.38)	0.0	0.3 (0.03)	-1.3 (0.71)	0.0	0.3 (0.04)	0.4 (1.07)	0.0
非製造	0.8 (0.01)	1.0 (0.24)	0.0	1.7 (0.02)	1.9 (0.34)	0.0	2.1 (0.02)	1.7 (0.56)	0.0

(注)  $\mu_{U_P}$  は真の意味で「分かる」企業の反実仮想的な「物価見通し」の平均値、 $\mu_P$  は「分かる」企業の「物価見通し」の平均値、 $B$  は無回答バイアスの試算値を意味する。  
( ) 内は標準誤差。

しかしながら、欠測が実際に生じたデータにもとづいて欠測メカニズムの仮定の妥当性を検証することは、多くの場合、不可能である。すなわち、ある企業の回答に欠測が生じた場合、それが回答そのものに依存したのか、回答そのものではなくその企業が有する何らかの属性に依存したのかを識別することは不可能である。それを識別するためには、欠測したデータを後日改めて徴求する必要がある。統計作成当局では、そうした作業は通常行われない。このため、欠測してしまった「真の値」がどのようなメカニズムによって欠測したのかを事後的に検証する手段は存在しない<sup>12</sup>。したがって、欠測メカニズムに関する仮定の妥当性は、ある程度、主観的な判断に委ねられることになる。その仮定が十分に説得的であるとみなせるならば、その仮定にもとづいて開発した欠測値補完の手法を統計作成に利用すべきである。

この点、本稿の議論は、5節でみたように、「分からない」という回答が多義的であること、そして、「分かる」という回答が多義的でないこと、というふたつの仮定に立脚している。6節のイベント・スタディの結果は、前者の仮定を支持する一方で、後者の仮定の妥当性については何も教えてくれない。したがって、「分かる」という回答も「分からない」という回答も、ともに多義的であるという仮定を置くことにより、バイアスに関する議論をすることもまた可能で意義があると考えられる。現時点では、本稿の仮定がこうした他の仮定に比べて、より説得的であるとは必ずしも言えない。その意味で、本稿が提示した方法によって欠測値補完を行うことは時期尚早であると考えられる。今後、さまざまな分析を蓄積することによって、どのような仮定が妥当か、議論を深めていくことが必要である。

## 10 まとめ

本稿では、短観の「物価見通し」における「分からない」という回答の中から、真の意味では「分かる」と推察される回答を機械学習の手法を用いて識別した。そのうえで、真の意味では「分かる」と識別された企業が有する反実仮想的な「物価見通し」を推定した。

分析の結果は、次の四点にまとめられる。第一に、真の意味では「分かる」が、何らかの理由で「分からない」と回答した企業は、きわめて少ない。逆にいえば、「分からない」と回答したほとんどすべての企業は、真の意味において「分からない」と考えられる。第二に、真の意味では「分かる」と識別された企業は、真の意味でも「分からない」と識別された企業に比べて、相対的に企業規模が小さく非製造業に多い。第三に、真の意味で

<sup>12</sup>宇都宮・園田 [1] は、欠測がランダムに生じることを検証している数少ない例である。こうした検証が可能なのは、短観における売上高や設備投資計画などの年度計画項目では、複数の調査回にわたって何度も同じ質問を行うことにより、補完すべき「真の値」が時間の経過とともに明らかになるためである。

は「分かる」と識別された企業の反実仮想的な「物価見通し」は、公表集計値と統計的に有意に異なる。第四に、以上を踏まえると、短観の「物価見通し」の無回答バイアスは、統計的には無視できる。

以 上

## 参考文献

- [1] 宇都宮浄人・園田桂子 (2001) 「「全国企業短期経済観測調査」における欠測値補完の検討」日本銀行ワーキングペーパー・シリーズ No.01-11.
- [2] 宇野洋輔・永沼早央梨・原尚子 (2017) 「企業のインフレ予想形成に関する新事実：PART I —粘着情報モデル再考—」日本銀行ワーキングペーパー・シリーズ No.17-J-3.
- [3] 土屋隆裕 (2010) 「調査への指向性変数を用いた調査不能バイアスの二段補正—「日本人の国民性第12次全国調査」への適用—」『統計数理』58, pp.25-38.
- [4] 統計推進会議 (2017) 「最終取りまとめ」 [HTTPS://WWW.KANTEI.GO.JP/JP/SINGI/TOUKEIKAIKAKU/PDF/SAISHU\\_HONBUN.PDF](https://www.kantei.go.jp/jp/singi/toukeikaikaku/pdf/saishu_honbun.pdf)
- [5] 内閣府経済社会総合研究所景気統計部 (2017) 「欠測値補完に関する調査研究報告書【詳細版】」 [HTTP://WWW.ESRI.CAO.GO.JP/JP/STAT/REPORT/REPORT\\_ALL\\_DETAIL.PDF](http://www.esri.cao.go.jp/jp/stat/report/report_all_detail.pdf)
- [6] 平川貴大・鳩貝淳一郎 (2012) 「ビジネスサーベイにおける欠測値補完の検討—全国企業短期経済観測調査(短観)のケース—」日本銀行ワーキングペーパー・シリーズ No.12-J-8.
- [7] 星野崇宏 (2009) 『調査観察データの統計科学：因果推論・選択バイアス・データ融合』岩波書店.
- [8] 星野崇宏 (2010) 「調査不能がある場合の標本調査におけるセミパラメトリック推定と感度分析：日本人の国民性調査データへの適用」『統計数理』58, pp.3-23.
- [9] BREIMAN, LEO (1996) “BAGGING PREDICTORS,” *Machine Learning* 24, pp.123-140.
- [10] ELKAN, CHARLES AND KEITH NOTO (2008) “LEARNING CLASSIFIERS FROM ONLY POSITIVE AND UNLABELED DATA,” *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*, pp.213-220.
- [11] FAWCETT, TOM (2006) “AN INTRODUCTION TO ROC ANALYSIS,” *Pattern Recognition Letters*, 27 pp.861-874.
- [12] HECKMAN, JAMES, J. (1979) “SAMPLE SELECTION BIAS AS A SPECIFICATION ERROR,” *Econometrica* 47, pp.153-161.

- [13] KHAN, SHEHROZ S. AND MICHAEL G. MADDEN (2014) “ONE-CLASS CLASSIFICATION: TAXONOMY OF STUDY AND REVIEW OF TECHNIQUE,” *The Knowledge Engineering Review* 29, PP.345-374.
- [14] LEE, WEE SUN AND BING LIU (2003) “LEARNING WITH POSITIVE AND UNLABELED EXAMPLES USING WEIGHTED LOGISTIC REGRESSION,” *Proceedings of the 20th international conference on machine learning*, PP.448-455.
- [15] LI, WENKAI, GUO, QINGHUA, AND CHARLES ELKAN (2011) “A POSITIVE AND UNLABELED LEARNING ALGORITHM FOR ONE CLASS CLASSIFICATION OF REMOTE-SENSING DATA,” *IEEE Transactions on Geoscience and Remote Sensing* 49, PP.717-725.
- [16] LITTLE, RODERICK J. A. AND DONALD B. RUBIN (2002) *Statistical Analysis with Missing Data, 2nd Edition*, HOBOKEN, N.J., WILEY.
- [17] LIU, BING, LEE, WEE SUN, YU, PHILIP S., AND XIAOLI LI (2002) “PARTIALLY SUPERVISED CLASSIFICATION OF TEXT DOCUMENTS,” *Proceedings of the 19th international conference on machine learning*, PP.387-394.
- [18] LIU, BING, DAI, YANG, LI, XIAOLI, LEE, WEE SUN, AND PHILIP S. YU (2003) “BUILDING TEXT CLASSIFIERS USING POSITIVE AND UNLABELED EXAMPLES,” *Proceedings of the third IEEE international conference on data mining*, PP.179-186.
- [19] MORDELET, FANTINE AND JEAN-PHILIPPE VERT (2014) “A BAGGING SVM TO LEARN FROM POSITIVE AND UNLABELED EXAMPLES,” *Pattern Recognition Letters* 37, PP.201-209.
- [20] TAX, DAVID M. J. AND ROBERT P. W. DUIN (1999) “SUPPORT VECTOR DOMAIN DESCRIPTION,” *Pattern Recognition Letters* 20, PP.1191-1199.
- [21] TAX, DAVID M. J. AND ROBERT P. W. DUIN (2004) “SUPPORT VECTOR DATA DESCRIPTION,” *Machine Learning* 54, PP.45-66.
- [22] YANG, PENGYI, LIU, WEI, AND JEAN YANG (2017) “POSITIVE UNLABELED LEARNING VIA WRAPPER-BASED ADAPTIVE SAMPLING,” *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, PP.3273-3279.

## A 「物価全般」における選択肢(11)(12)(13)

3節でみたように、「物価全般」には、(11)先行きについては不確実性が大きいから、(12)変動したとしても経営にほとんど影響がないため意識していないから、(13)その他、という三つの選択肢がある。本稿では、これら三つの選択肢を選んだ回答を一括りにして「分からない」と定義している。本補論では、こうした定義の妥当性について議論する。

まず、三つの選択肢を選んだ回答の分布を確認しておく。表 A.8 をみると、選択肢(12)(13)を選んだ回答が相対的に少なく、また、年限が長くなるほど、選択肢(11)の割合が高くなることが確認できる。これを踏まえ、仮に三つの選択肢が異なる情報を有しているとすれば、三つの選択肢を一括りにすることによって、とりわけ年限の長い見直しにおいて、選択肢(12)と(13)の情報が薄められることになる。

表 A.8: 選択肢(11)(12)(13)を選んだ回答の分布

	1年後		3年後		5年後	
選択肢(11)	19218	(0.80)	42208	(0.89)	58293	(0.91)
選択肢(12)	3414	(0.14)	3781	(0.08)	3965	(0.06)
選択肢(13)	1326	(0.06)	1685	(0.04)	2009	(0.03)

(注) 表中の () 内は、割合。

以下では、本稿の分析の枠組みに即して、三つの選択肢が異なる情報を有しているかを議論する<sup>13</sup>。まず、6節のイベント・スタディに即してみよう。6節では、選択肢(11)(12)(13)のいずれかを選んだ回答を「分からない」と定義したが、ここでは、選択肢(12)と(13)のそれぞれの選択肢を選んだ回答のみを「分からない」と定義し、同様のイベント・スタディを行う。もし、6.2節で観察されたような統計的に有意な変化が観察できなくなるのであれば、選択肢(12)と(13)が(11)とは異なる情報を有していると判断できる。

表 A.9 をみると、選択肢(12)と(13)のいずれを「分からない」と定義した場合でも、すべての年限において、窓口変更時に統計的に有意な変化が確認できる。すなわち、選択肢(12)と(13)においても、回答の「揺れ」が観察される<sup>14</sup>。このことは、選択肢(12)と(13)のいずれを「分からない」と定義した場合でも、(1)式の仮定が満たされることを意味している。

次に、本稿のアルゴリズムの中でロジスティック回帰モデルを用いることを踏まえ、以下のような実験を行う。まず、「分かる」という回答を1,000個をランダムにサンプリング

<sup>13</sup>議論の前提として、本稿では、三つの選択肢が異なる情報を有しているかどうかは、先験的に明らかでなく、定量的な分析結果にもとづいて判断されるべきであると考えている。

<sup>14</sup>仔細にみれば、状態遷移確率の差は、いずれの年限でみても、選択肢(12)の方が選択肢(13)より大きい。すなわち、選択肢(12)の方が回答の「揺れ」が大きい。

表 A.9: 「分かる」から「分からない」への状態遷移確率

分からない=選択肢 (12)			
	1 年後	3 年後	5 年後
窓口変更時 ( $z = 1$ )	0.0098 (0.0065,0.0134)	0.0139 (0.0095,0.0186)	0.0156 (0.0110,0.0211)
窓口不変時 ( $z = 0$ )	0.0027 (0.0024,0.0029)	0.0036 (0.0032,0.0039)	0.0043 (0.0039,0.0048)
差	0.0071	0.0103	0.0113
分からない=選択肢 (13)			
	1 年後	3 年後	5 年後
窓口変更時 ( $z = 1$ )	0.0030 (0.0013,0.0050)	0.0057 (0.0028,0.0085)	0.0071 (0.0038,0.0108)
窓口不変時 ( $z = 0$ )	0.0010 (0.0008,0.0012)	0.0014 (0.0012,0.0017)	0.0025 (0.0022,0.0028)
差	0.0020	0.0042	0.0046

(注) 表中の () 内は、ブートストラップ法 (パーセンタイル法、窓口変更時は反復 2000 回、窓口不変時は反復 200 回) によって計算した 95%信頼区間を示す。

して固定する。次に、「分からない」という回答については、三つの選択肢 (11)(12)(13) のそれぞれについて、ランダムに 1,000 個の回答をサンプリングする。そのうえで、各年限について、「分からない」の「ラベル」だけが異なる、三通りのロジスティック回帰モデル (サンプルサイズ 2,000) を推定し<sup>15</sup>、標本内予測精度を比較する<sup>16</sup>。もし、三通りのロジスティック回帰モデルの標本内予測精度が統計的に有意に異なるのであれば、三つの選択肢がそれぞれに異なる情報を有していると判断できる。

表 A.10 は、「分からない」の「ラベル」だけが異なる、三通りのロジスティック回帰モデルの標本内予測精度を示している。この結果から、次の三点を指摘することができる。第一に、三つの選択肢のいずれを用いたとしても、すべての年限において、AUC は 0.7 を上回る。すなわち、三つの選択肢のいずれを「分からない」とした場合でも、予測精度は相応に高い。第二に、1 年後の選択肢 (11) と (12) を除き、三つの選択肢の予測精度は、統計的に有意に異なる。いずれの年限でみても、選択肢 (11) を「分からない」とした場合の予測精度が最も低い一方、選択肢 (13) を「分からない」とした場合の予測精度が最も高い。このことは、直観的にいえば、選択肢 (11) より選択肢 (13) の方が分類が容易であることを意味する。第三に、選択肢 (13) を「分からない」と定義した場合、AUC は 0.804~0.857

<sup>15</sup> 説明変数は、各年限とも、7.4 節で示した変数のセットから DI の「先行き」を除いたものとした。

<sup>16</sup> 予測精度の比較には、機械学習の分野で広く用いられている、AUC と呼ばれる統計量を用いる。AUC は 0.5 と 1 の間の値をとり、1 に近いほどモデルの予測精度が高いと評価する。詳細は、Fawcett [11] を参照。

表 A.10: ロジスティック回帰モデルの標本内予測精度

「分からない」ラベル	1年後	3年後	5年後
	AUC		
選択肢 (11)	0.749	0.708	0.711
選択肢 (12)	0.770	0.790	0.765
選択肢 (13)	0.857	0.825	0.804
	AUC の差の有意性 ( $p$ 値)		
選択肢 (11) vs 選択肢 (12)	0.132	0.000	0.000
選択肢 (11) vs 選択肢 (13)	0.000	0.000	0.000
選択肢 (12) vs 選択肢 (13)	0.000	0.003	0.002

(注)  $p$  値は、ブートストラップ法 (パーセンタイル法、反復 2000 回) を用いて計算した。

である。すなわち、選択肢 (13) を「分からない」と定義した場合の予測精度が相対的に高いとはいえ、選択肢 (11) との差は、いずれの年限でも、たかだか 0.1 程度である。

以上の分析結果をまとめると、次のとおりである。同一条件下での予測精度という面からみれば、三つの選択肢は、いくらか異なる情報を有しているとみられる。もっとも、(1) 同一条件下での予測精度の差が AUC でみて 0.1 程度にとどまること、(2) 表 A.9 のイベント・スタディの結果から、三つの選択肢の違いを観察できないこと、という二点から、三つの選択肢の差異は限定的であると判断できる。これを踏まえ、本稿では、三つの選択肢を一括りにして扱う。

## B Rのコード

7.2節のアルゴリズムは、以下のRのコードで実行した。

```
library(dplyr)
library(kernlab)

data_p <- data %>%
  filter(wakaruyo == 1)
data_u <- data %>%
  filter(wakaruyo == 0)

val_names <- colnames(data)
f <- as.formula(paste("wakaruyo~",
  paste(val_names[val_names != "wakaruyo"],
    collapse = "+")))

# step 1 #
sigma <- 0.05
outlier <- 0.02
M <- length(data_p$wakaruyo)
occ_svm <- ksvm(x = f, data = data_p, type = "one-svc",
  kernel = "rbfdot", kpar = list(sigma = sigma),
  C = 1/(outlier*M),
  nu = outlier)
occ_pred <- predict(occ_svm, newdata = data_u)
L <- length(occ_pred) - sum(occ_pred)

# step 2 #
new_positive <- occ_pred
data_u_temp_U <- data_u
dif_U <- 1

repeat{
  if(dif_U > 0){
    U_size_temp <- length(data_u_temp_U$wakaruyo)
    data_u_temp_N <- data_u_temp_U %>%
      cbind(new_positive) %>%
      filter(new_positive == FALSE) %>%
      select(-new_positive)
    data_u_temp_U <- data_u_temp_U %>%
      cbind(new_positive) %>%
      filter(new_positive == TRUE) %>%
      select(-new_positive)

    Z <- sample(nrow(data_p), size = L)
    glm_fit <- glm(formula = f,
      data = rbind(data_p[Z,], data_u_temp_N),
      family = binomial)
    pred <- predict(glm_fit, newdata = data_u_temp_U, type = "response")
    new_positive <- as.data.frame(pred$pred >= 0.5)
    colnames(new_positive) <- c("new_positive")
    dif_U <- U_size_temp - length(data_u_temp_U$wakaruyo)
  } else {
    break
  }
}
```