연구용역 결과 보고서

체납 정리기간 결정요인에 관한 연구

위 연구용역에 관하여 결과 보고서를 제출합니다.

2018년 12월 20일

책임연구원 박상수 (고려대학교 경제학과 교수)

책임연구원 한치록 (고려대학교 경제학과 교수)

차 례

1 도입 **1**

1.1 연구의 목적과 내용 1

1.2 연구 방법 3

1.3 해외 사례 3

2 분석 자료 구축 **7**

2.1 개요 7

2.2 국세청 제공 자료 8

2.3 자료 구축 시 특이 사항 11

2.4 자료 생성 12

3 체납 결정요인 분석 **19**

3.1 모형과 방법론 19

3.2 데이터 및 요약통계량 25

3.3 추정 결과 25

3.4 모형과 추정법의 수학적 설명 30

3.4.1 변수의 설명과 표기 30

3.4.2 2단 모형(Two-Tier Model) 34

3.4.3 단일 모형(One-Tier Model) 37

3.4.4 체납 확률, 체납 기간, 성실 납부 성향의 추정 42

3.4.5 모형 추정의 세부 내용 : 문턱값과 추정 모형의 선정 43

3.4.6 모형 추정 결과 51

3.4.7 부록 A: 모형별 추정 결과 52
3.4.8 부록 B: 표준화에 이용한 a와 b 및 변수들의 평균 62

4 결론 **66**

표 차례

2.1 납세고지내역 데이터 구조 9

2.2 수납내역 데이터 구조 10

2.3 개인속성 데이터 구조 10

2.4 징수유예신청기본 데이터 구조 11

3.1 12개 모형 비교 20

3.2 모형별 종합적 예측 성과 26

3.3 M6B 모형의 추정 결과 27

3.4 모형의 구분 41

3.5 모형별 체납 확률과 체납될 때 평균 체납 기간 42

3.6 모형별 납세 의무자의 성실 납부 성향 점수 43

3.7 체납확률과 체납 시 체납기간 기댓값의 추정 44

3.8 선정된 문턱값과 해당 문턱값에서의 $FP^2 + FN^2$ 48

3.9 모형별 평균 체납기간 RMSE 49

3.10 모형별 평균 종합적 예측 성과 50

3.11 M1B 추정 결과 52

3.12 M2B 추정 결과 54

3.13 M3B 추정 결과 56

3.14 M4B 추정 결과 58

3.15 M5B, M6B 추정 결과 60

3.16 분석에 사용된 데이터 전체의 평균과 a, b 62

3.17 체납인 경우와 미체납인 경우의 평균 64

그림 차례

3.1 문턱값에 따른 체납자와 비체납자의 올바른 예측 확률(M6B 모형) 22
3.2 문턱값에 따른 모형별 오류비율과 그 제곱합 45

제 1 장

도입

1.1 연구의 목적과 내용

□ 본 연구에서는 과세액을 기한 내에 완납하지 못하고 체납될 확률과 체납이 발생하는 경우 체납 기간을 추정하는 모형을 제시하고자 함

○ 모형을 통해 예측되는 체납확률과 체납기간으로 납부자 성향을 파악하는 것을 목적으로 함

○ 이를 위하여 국세청 기초자료로부터 분석용 데이터를 구축하고, 구축된 데이터를 실제 분석하는 것을 목표로 함

□ 연구내용 I: 데이터 구축

○ 주요 연구내용 중 하나는 국세청 기초자료로부터 분석 가능한 데이터를 생성하는 것임

○ 국세청으로부터 제공 받은 기초 자료는 납세고지 자료, 수납 자료, 납세자 정보에 관한 자료, 징수유예 신청 경력에 관한 자료, 사업체 정보에 관한 자료 등으로 구성되어 있으며, 분석에 맞도록 가공된 데이터가 아니므로 가공이 필요함

› 특히 납세 고지서별 완납까지 걸린 시간과 체납여부 등을 알기 위해서는 수납 자료와 연결하여야 하나 여기에는 상당한 노력이 필요함

○ 이러한 가공된 데이터를 국세청과 같이 별도의 본연의 업무를 가진 기관에서 작성하고 제공하는 데에는 인력과 시간에 있어 상당한 비용이 소요됨

› 분석의 목적에 맞추어 데이터를 가공하여 추출하는 것은 전문성을 요구하는 작업으로, 특화된 연구진이 아닌 일반 업무 부서에서 처리하는 데에는 상당한 인력과 시간이 요구됨

○ 본 연구에서는 기초자료 제공기관(국세청)과 연구진의 업무를 효율적으로 분담하는 모델을 제시함

○ 기초자료 제공기관은 연구의 목적이나 방법론을 고려한 특화된 데이터를 생성하기보다는 추가 업무 부담을 최소화한 상태에서 자료를 추출할 수 있도록 하고, 연구진은 이러한 자료로부터 전문성을 활용하여 필요한 정보를 추출할 수 있도록 함으로써, 추후 신규 연구나 기존 연구의 확장 시 효율성을 극대화할 수 있도록 하였음

○ 본 장에서는 기초자료 제공기관에서 통상적인 방식으로 추출한 자료를 효율적인 방법으로 가공하여 다양한 연구 목적에 맞도록 2차 자료를 생성하는 방법을 제시함

□ 연구내용 II: 체납확률 및 체납 시 정리기간 결정요인 분석

○ 본 연구를 위하여 국세청으로부터 제공된 데이터(납세고지에 관한 정보, 수납이력에 관한 정보, 납세자 체납 이력에 관한 정보, 사업 정보 등)를 활용하여 체납될 확률과 체납이 발생하는 경우 체납 정리기간을 추정하는 모형을 제시하고 분석함

○ 본 연구를 위하여 국세청이 선별하여 제공한 자료만을 활용하였음

○ 모형 추정에 이용한 데이터는 납세액이 5백만 원 이상인 납세의무자들의 2015년–2017년 데이터

□ 연구의 구성

○ 제2장에서는 국세청 제공 기초 자료로부터 분석가능한 형태의 자료를 생성하는 방법을 설명함

○ 제3장에서는 제2장에서 변환한 자료를 바탕으로 체납확률 및 체납 정리기간 결정요인들을 분석함

○ 제4장은 보고서를 맺음

□ 정책적 제언: 연구 결과를 바탕으로 개별 납세자의 체납 가능성과 체납 시 정리기간을 예측하는 데에 활용할 수 있음

○ 본 연구는 국세청 데이터의 일부만을 사용하여 진행되었으며, 연구 결과는 향후 체납 가능성과 체납 시 정리기간 예측을 위한 데이터 기반 모형 개발에 지침으로 활용될 수 있음

1.2 연구 방법

□ 국세청 기초 자료는 크게 납세고지 자료, 수납 자료, 납세자 정보에 관한 자료, 징수유예 신청 경력에 관한 자료, 사업체 정보에 관한 자료 등으로 구성되어 있음

□ 기초 자료를 가공하여 각 납세고지내역별로 체납여부, 완납일까지 소요 기간, 체납기간을 만들어 내고, 여기에 개인의 속성, 사업체의 속성 등을 추가한 자료를 생성

□ 작성된 자료를 바탕으로, 체납여부에 관한 이항반응모형, 납부기간에 관한 “duration” 모형 등을 이용하여 체납 정리 행태를 분석함

□ 상세한 연구 방법은 각각 2장과 3장에서 설명함

1.3 해외 사례

□ 본 소절에서는 체납과 관련하여 데이터 분석모델을 활용하는 해외 사례들을 국가별로 간략히 요약함[1]

□ 호주

---

[1]본 소절의 내용은 OECD (2014) “Working Smarter in Tax Debt management”의 내용과 빅데이터와 조세행정의 최근 해외 트렌드에 대한 한국조세재정연구원(2018)의 연구 내용을 발췌하여 간략히 요약한 것임. 주요국의 체납정리 인프라 현황(한국조세연구원, 2010)도 참조.

○ 호주는 2011년부터 측정 데이터에 기반하여 조세체납액을 징수하는 'Debt Right Now (DRN) 프로그램'을 도입함

› 납세자의 지급능력을 분석하는 모델과 지급성향을 분석하는 모델로 구성됨

› 과거 납세자 납부성향 및 재무현황을 파악하여 위험점수를 산정하고 점수가 높을 수록 강한 제재수단을 적용

○ DRN 프로그램 도입에도 불구하고 도입 이후 2년간 체납액은 증가

○ 호주 국세청은 납세자의 체납발생이유 또는 체납액의 계속적 증가 여부 등 기타 다양한 요인도 고려하여 납세자에 맞는 체납 정리 방안을 적용하는 모형으로 변경하고자 함(이상, 한국조세재정연구원, 2018)

› 주요 대학과 협업하여 추진 중으로 예측과 사전처방적 분석으로 전환하고 있음

□ 네덜란드

○ 네덜란드는 납세액 징수 단계별로 납세자의 납부 기록 및 현황을 한번에 파악할 수 있는 시스템을 운영중임

› 체납 시 납세고지문 발송, 압류명령, 압류, 경매와 같은 실질적인 법적절차 시행 등 어느 단계에서 체납이 정리되었는지 파악이 가능함

○ 과거 축적된 데이터를 통해 과세당국은 향후 납세자의 납부성향을 예측하고, 어떤 방법이 체납 정리에 가장 효과적이었는지 파악하는 데 활용하며, 또한 실시간 분석을 통해 납세자에게 알맞은 징수방안을 적용함(이상, 한국조세재정연구원, 2018)

□ 벨기에

○ 벨기에 국세청은 채무불이행 위험 툴을 이용하며 위험모델을 정교화

○ 위험분석툴은 지불능력, 유동성, 수익성 등 몇 가지 요인을 사용하여 지급불능을 예측하는 모델

□ 캐나다

○ 캐나다 국세청은 납세자 특성과 과거 행동이력에 기초하여 사안별 중요도를 분별하여 상이한 조치를 취하는 위험기초접근법을 적용하는 자동화된 전략을 이용함

○ 개별납세자의 행동에 기초하여 위험점수를 산정하는 정교한 예측 모델을 사용함으로써 직관적인 의사결정에서 증거기반 접근법으로 이동

□ 영국

○ 영국 국세청은 통합 체납관리시스템을 가지고 있음

○ IT 시스템은 한 납세자가 지고 있는 주요 의무의 모든 부채정보와 세금공제까지 징수관에게 보여 줌

○ 군집툴을 사용하여 납세자들을 분류함

› 납세자들은 과거 납세이력, 납부능력 등에 기초하여 군집 분류되고 맞춤형 조치를 받음

› 위험분류 엔진은 각 납세자를 평가하여 체납환수 활동이 가장 효과적이 되도록 목표화하여 사용됨

□ 스페인

○ 스페인에서는 모든 납세자를 납부능력, 체납액 규모, 체납액의 변동성 세 가지 기준으로 분류함

› 위험 프로필은 중간, 높음, 매우 높음으로 구분됨

› 예측 분석은 청산 가능성이 높은 자에게 경고알림 메시지를 제공함

○ 2007년 데이터마이닝 시스템 시작

○ 실제 미납자에 대한 위험 평가는 사안을 다루는 담당자에게는 접근 불가

○ 일선 담당자들은 예를 들어 파산신청이나 성실납세 철회 같은 일반적인 툴박스를 사용하도록 함

□ 아일랜드

○ 아일랜드 국세청은 납세자의 특성과 과거행동이력에 기초하여 목표 납세자를 제대로 정하는 조치 전략을 가지고 있음

○ 체납액은 위험에 근거하여 우선 순위를 정하고 차별화

○ 통합 세무기록 및 사안별 관리시스템을 가지고 있음

› 모든 납세자의 기본 데이터와 부채, 징수, 주요 세목의 대차가 하나의 기록 시스템에 있어 납세자의 전반적인 관찰이 가능

○ 아일랜드는 청산 예측 모델과 체납자 체납징수를 위치 맵핑하는 GIS를 개발 중

제 2 장

분석 자료 구축

2.1 개요

□ 본 장에서는 국세청 제공 기초자료로부터 체납기간 결정 요인 분석에 적합한 형태의 데이터를 생성하는 것을 목적으로 함

□ 국세청 기초자료는 국세업무 경험을 토대로 체납과 높은 연관을 가질 것으로 예상되는 변수들을 위주로 본 연구진에게 전달되었음

○ 변수의 상세 내역과 데이터 구조는 2.2절 참조

□ 국세청 담당자와 연구진의 각각의 장점을 이용한 효율적인 업무분담 모델을 이용하였음

○ 국세청으로부터 제공 받은 기초 자료는 납세고지 자료, 수납 자료, 납세자 정보에 관한 자료, 징수유예 신청 경력에 관한 자료, 사업체 정보에 관한 자료 등으로 구성되어 있으며, 분석에 맞도록 가공된 데이터가 아니므로 가공이 필요함

› 특히 납세 고지서별 완납까지 걸린 시간과 체납여부 등을 알기 위해서는 수납 자료와 연결하여야 하나 여기에는 상당한 노력이 필요함

○ 이러한 가공된 데이터를 국세청과 같이 별도의 본연의 업무를 가진 기관에서 작성하고 제공하는 데에는 인력과 시간에 있어 상당한 비용이 소요됨

› 분석의 목적에 맞추어 데이터를 가공하여 추출하는 것은 전문성을 요구하는 작업

으로, 특화된 연구진이 아닌 일반 업무 부서에서 처리하는 데에는 상당한 인력과 시간이 요구됨

○ 본 연구에서는 기초자료 제공기관(국세청)과 연구진의 업무를 효율적으로 분담하는 모델을 제시함

○ 기초자료 제공기관은 연구의 목적이나 방법론을 고려한 특화된 데이터를 생성하기보다는 추가 업무 부담을 최소화한 상태에서 자료를 추출할 수 있도록 하고, 연구진은 이러한 자료로부터 전문성을 활용하여 필요한 정보를 추출할 수 있도록 함으로써, 추후 신규 연구나 기존 연구의 확장 시 효율성을 극대화할 수 있도록 하였음

○ 본 장에서는 기초자료 제공기관에서 통상적인 방식으로 추출한 자료를 효율적인 방법으로 가공하여 다양한 연구 목적에 맞도록 2차 자료를 생성하는 방법을 제시함

2.2 국세청 제공 자료

□ 국세청 제공 기초자료는 개인으로서 복식부기의무자 중 임의 선정한 자 6만여 명 자료로서, 다음 구조를 가짐

① 납세고지내역(파일명 : GOJI.CSV)

〈표 2.1〉 납세고지내역 데이터 구조

칼럼명	한글명
TIN	납세자통합관리번호
LVY_DCS_ID	징수결정ID
LVY_DCS_TXPR_BRKD_SN	징수결정납세자내역일련번호
BMAN_TIN	사업자납세자통합관리번호
LVY_CHRG_TXHF_OGZ_CD	징수결정청서
NTCP_ISU_DT	고지일자
PMT_DDT	최종납부기한
TXAMT1	차감고지세액
AFT_YN	체납여부
TXP_DUTY_TMNT_RSN_CD	납세의무종결사유코드
FRW_CNT	고지서발송횟수
RPCL_CNT	징수유예횟수
DPAT_RSTP_CNT	체납처분유예횟수

› 각 납세고지별로 징수결정ID, 고지일자, 최종납부기한, 차감고지세액을 비롯한 고지내역에 관한 정보가 있으며, 2018년 10월 현재 체납여부, 납세의무종결사유코드에 관한 정보가 추가적으로 있음

› 본 데이터에는 납세고지 내역만 존재하며, 이를 이하 ②의 수납내역과 연결시켜 최종납부일과 기한내 납부 여부에 관한 변수를 생성시킬 것임(참고로, '체납여부' 변수는 데이터가 추출된 2018년 10월 현재 시점의 체납 여부를 나타내므로 분석에 사용할 수 없으나, 구축된 자료의 타당성 여부를 점검하는 데에 간접적으로 참고하였음)

› 납세자에 관한 정보와 연결시킬 수 있도록 해 주는 납세자통합관리번호가 있어 이를 이하 ③의 개인속성 및 ④의 징수유예신청 내역과 결합시킬 것임

② 수납내역(파일명 : SUNAB.CSV)

〈표 2.2〉 수납내역 데이터 구조

칼럼명	한글명
TIN	납세자통합관리번호
LVY_DCS_ID	징수결정ID
LVY_DCS_TXPR_BRKD_SN	징수결정납세자내역일련번호
BMAN_TIN	사업자납세자통합관리번호
ROM_ID	수납ID
ROM_DT	수납일자
ROM_AMT	수납금액

› 수납내역 기초자료는 수납 내역을 정리한 것이며, 이 기초자료를 처리하고 앞의 ① 고지내역과 결합하여 고지서별 체납여부와 최종납부일을 찾고, 최종납부일과 ①의 '최종납부기한'을 비교하여 체납 여부와 체납 정리기간을 파악할 것임

③ 개인속성(파일명 : PERSON.CSV)

〈표 2.3〉 개인속성 데이터 구조

칼럼명	한글명
TIN	납세자통합관리번호
AGE	연령
SEX_CD	성별코드
MFH_CNT	세대원수
JONGSO_2015	2015년종합소득금액
JONGSO_2016	2016년종합소득금액
JONGSO_2017	2017년종합소득금액
CAR_YN	자동차보유여부
NE_CMFG_YN	명의위장사업자여부

④ 징수유예신청기본(파일명 : TTNANEMA04.CSV)

〈표 2.4〉 징수유예신청기본 데이터 구조

칼럼명	한글명
TIN	납세자통합관리번호
RPCL_APLN_CVA_ID	징수유예신청민원ID
RPCL_CL_CD	징수유예구분코드
RPCL_APLCD	징수유예신청구분코드
RPCL_KND_CD	징수유예종류코드
RPCL_RSN_CD	징수유예사유코드
RPCL_STRT_DT	징수유예시작일자
RPCL_END_DT	징수유예종료일자
RPCL_APRV_DT	징수유예승인일자
RPCL_CNCL_RSN_CD	징수유예취소사유코드
RPCL_PRGR_STAT_CD	징수유예진행상태코드

⑤ 재무제표

2.3 자료 구축 시 특이 사항

□ 자료 구축과 분석 시 보안을 가장 중요한 요소로 고려하였음

○ 특히 통계자료는 납세자의 과세정보를 직접적 방법 또는 간접적인 방법으로 확인할 수 없도록 작성되어야 함(국세기본법 제85조의 6)에 유의하고, 납세 정보가 국세청 서버 이외의 장소에 저장되지 않도록 보안에 만전을 기하였음

○ 이와 더불어 모든 납세 정보가 익명화 처리되고, 본 연구를 위하여 징수결정ID나 납세자통합관리번호 등이 무작위로 생성되어 개별 납세건을 식별할 수 없는 높은 보안 수준을 유지하였음

□ 보안을 위하여 R 서버를 이용하여 서울지방국세청 단말기로부터 작업을 하였음

○ 데이터는 국세청 본청 서버에 저장되어 있고, R 프로그램 또한 본청 서버에서 실행됨

○ 특정 단말기만을 사용하며 데이터 파일은 어떠한 형태로든 반출이 원천적으로 차단됨

□ 이러한 보안상의 고려는 불가피하게 연구진의 분석에 상당한 제약 요인으로 작용함

□ 본 연구에서는 서울청 해당 데이터만 이용하여 데이터 크기가 제한됨에도 불구하고 자료의 양이 막대하여 자료 구축의 효율성을 높이는 것이 중요하였음

2.4 자료 생성

□ 개요

○ 방대한 데이터 크기로 인하여 데이터 처리에 상당한 시간이 소요되므로, 여러 단계로 나누어 처리하여 효율성을 제고할 필요가 있음

○ 특히, 기초자료의 특성상 데이터에 불가피하게 입력오류가 포함되어 있을 수 있으므로 수많은 오류검증이 필요하며, 이를 위해서도 단계를 구분하는 것이 바람직함

□ (1단계) CSV 파일을 R이 직접 처리할 수 있는 rds 파일의 형태로 변형하여 저장함: R은 CSV 파일을 효율적으로 처리할 수 있음에도, 데이터의 크기 및 디버깅 필요로 인하여 매번 데이터를 읽어들이는 데에는 막대한 시간이 소요되므로 본 과정이 유용함

① 고지내역은 다음 코드로써 변환함(변환 후 GOJI.rds 파일 생성)

```
`%+%` <- paste0                                  # 코딩 간편화
`%포함%` <- function(x,y) x[grep(y,x)]           # 코딩 간편화

z <- read.csv('GOJI.csv', colClasses = 'character',
              stringsAsFactors = FALSE)
names(z) <- c('납세자통합관리번호', '징수결정ID',
              '징수결정납세자내역일련번호', '사업자납세자통합관리번호',
              '징수결정청서', '고지일자', '최종납부기한',
              '차감고지세액', '체납여부', '납세의무종결사유코드',
              '고지서발송횟수', '징수유예횟수', '체납처분유예횟수')
for (v in c('고지일자', '최종납부기한'))
    z[[v]] <- as.Date(z[[v]], format='%Y%m%d')
for (v in c('차감고지세액','고지서발송횟수','징수유예횟수',
            '체납처분유예횟수'))
    z[[v]] <- as.numeric(z[[v]])
saveRDS(z, 'GOJI.rds')
```

› 참고로, 변수 중 크기가 매우 큰 정수로 해석될 수 있는 값들이 포함되어 있어 R의 자동변환 기능을 활용할 경우 불필요하게 긴 시간이 소요되고 변환이 부정확하게 될 수 있으므로, 읽어들일 변수들의 자료형을 미리 지정하는 것(colClasses 옵션 참조)이 효율성을 높이는 방법임

› 변수명을 한글로 바꾸어 놓는 것이 직관적인 디버깅에 도움이 됨

② 수납내역은 다음 코드로써 변환함(변환 후 SUNAB.rds 파일 생성)

```
z <- read.csv('SUNAB.csv', colClasses = 'character',
              stringsAsFactors = FALSE)
names(z) <- c('납세자통합관리번호', '징수결정ID',
              '징수결정납세자내역일련번호', '사업자납세자통합관리번호',
              '수납ID', '수납일자', '수납금액')
for (v in c('수납일자')) z[[v]] <- as.Date(z[[v]], format='%Y%m%d')
for (v in c('수납금액')) z[[v]] <- as.numeric(z[[v]])
saveRDS(z, 'SUNAB.rds')
```

③ 개인속성은 다음 코드로써 변환함(변환 후 PERSON.rds 파일 생성)

```
z <- read.csv('PERSON.csv',
              colClasses = c('character','integer','character',
                'integer','numeric','numeric','numeric',NA,NA),
              stringsAsFactors = FALSE)
names(z) <- c('납세자통합관리번호', '연령', '성별코드', '세대원수',
              '종합소득2015', '종합소득2016', '종합소득2017',
              '자동차보유여부', '명의위장사업자여부')
saveRDS(z, 'PERSON.rds')
```

④ 징수유예 신청내역은 다음 코드로써 변환함(변환 후 TTNANEMA04.rds 파일 생성)

```
z <- read.csv('TTNANEMA04.csv',
              colClasses = c('character','character',NA,NA,NA,
                NA,'character','character','character',NA,NA),
              stringsAsFactors = FALSE)
names(z) <- c('납세자통합관리번호', '징수유예' %+%
              c('신청민원ID', '구분코드', '신청구분코드', '종류코드',
                '사유코드', '시작일자', '종료일자', '승인일자',
                '취소사유코드', '진행상태코드'))
for (v in names(z) %포함% '일자$')
    z[[v]] <- as.Date(z[[v]], format='%Y%m%d')
```

```
saveRDS(z, 'TTNANEMA04.rds')
```

⑤ 재무제표는 다음 코드로써 변환함(변환 후 재무제표.rds 파일 생성)

```
bs <- read.csv('재무제표.csv',
               colClasses = c(NA, 'character', 'character',
                 rep(NA, 440-3)),
               stringsAsFactors = FALSE)
names(bs)[2:3] <- c('사업자납세자통합관리번호', '납세자통합관리번호')
saveRDS(bs, '재무제표.rds', compress = FALSE)
```

□ (2단계) ②~④의 데이터를 ① 고지내역 데이터에 분석 가능한 형태로 추가함

○ 고지내역 데이터 ①에서 중복된 징수결정ID를 삭제하고, 데이터상으로 명백한 불일치가 있는 관측치(고지일자가 최종납부기한 이후인 것)와 차감고지세액이 0 이하인 관측치를 제외하고, 납세의무종결사유코드가 01, 08, ZZ인 것만 분석에 이용함

```
z <- readRDS('GOJI.rds')
z <- z[ave(z$징수결정ID, z$징수결정ID, FUN=length)==1,]  # 중복 제외
z <- z[z$고지일자 < z$최종납부기한,]
z <- z[z$차감고지세액>0,]
z <- z[z$납세의무종결사유코드 %in% c('01','08','ZZ'),]
```

○ 수납내역 데이터 ②에서 징수결정ID별로 수납횟수, 수납총액, 최초수납일, 최종수납일을 구하고 이를 고지내역 데이터 ①에 변수로 추가함. 이때 완납이면서 수납 기록이 없는 징수결정ID를 삭제함

```
uniqueLength <- function(x) length(unique(x))
myTapply <- function(v, FUN) tapply(SN[[v]], SN$id.local, FUN)
first <- function(x) x[1]

SN <- readRDS('SUNAB.rds')
SN$id.local <- factor(SN$징수결정ID)
w <- data.frame(징수결정ID = myTapply('징수결정ID', first))
w$수납횟수 <- myTapply('수납금액', length)
w$수납총액 <- myTapply('수납금액', sum)
w$최초수납일 <- myTapply('수납일자', min)
w$최종수납일 <- myTapply('수납일자', max)
for (v in c('최초수납일', '최종수납일'))
```

```
    w[[v]] <- as.Date(w[[v]], origin = '1970-01-01')
idx <- match(z$징수결정ID, w$징수결정ID)
for (v in setdiff(names(w), c('징수결정ID', 'id.local')))
    z[[v]] <- c(w[idx, v])
z <- z[!(z$납세의무종결사유코드=='01' & is.na(z$최종수납일)),]
```

○ 납세자별 정보를 개인별 데이터 ③으로부터 추출하여 고지내역 데이터 ①에 변수로 추가함

```
w <- readRDS('PERSON.rds')
idx <- match(z$납세자통합관리번호, w$납세자통합관리번호)
for (v in setdiff(names(w), 'seqno')) z[[v]] <- w[idx,v]
```

○ 종속변수와 여타 관련 변수들을 생성함: 기한내 완납 여부, 체납 여부, 체납 시작일, 체납 종료일, 납세자별 연도별 고지서 개수. 이 중 납세자별 연도별 고지서 개수를 제외하고는 주 데이터에 추가함

```
z$기한내완납 <- as.numeric(z$납세의무종결사유코드=='01' &
    !is.na(z$최종수납일) & z$최종수납일 <= z$최종납부기한)
z$체납 <- as.numeric(z$기한내완납==0)
z$체납시작일 <- numToDate(ifelse(z$체납, z$최종납부기한, NA))
z$체납종료일 <- numToDate(ifelse(z$체납, z$최종수납일, NA))
z$체납종료일[z$체납 & z$납세의무종결사유코드=='ZZ'] <-
    as.Date('2018-10-09')

z$pid <- factor(z$납세자통합관리번호)
z$연도 <- as.numeric(format(z$고지일자, '%Y'))
p <- data.frame(납세자통합관리번호 =
                c(tapply(z$납세자통합관리번호, z$pid, first)))
for (j in unique(z$연도)) {
    v <- '고지서개수' %+% j
    p[[v]] <- c(tapply(as.numeric(z$연도==j), z$pid, sum))
}
z$pid <- NULL
```

○ 유예신청 데이터 ④로부터 납세자통합관리번호별로 연도별 유예신청 건수를 생성하고, 납세자통합관리번호별 정보(앞 단계에서 생성한 변수 포함)를 주 데이터에 변수로 추가함

```
y <- readRDS('TTNANEMA04.rds')
y$연도 <- as.numeric(format(y$징수유예시작일자, '%Y'))
y$pid <- factor(y$납세자통합관리번호)
u <- data.frame(납세자통합관리번호 =
    tapply(y$납세자통합관리번호, y$pid, first))
for (j in sort(unique(y$연도))) {
    v <- '유예신청건수' %+% j
    u[[v]] <- unlist(tapply(as.numeric(y$연도==j), y$pid, sum))
}
y$pid <- NULL

idx <- match(p$납세자통합관리번호, u$납세자통합관리번호)
for (v in setdiff(names(.u), c('납세자통합관리번호'))) {
    p[[v]] <- u[idx,v]
    p[[v]][is.na(p[[v]])] <- 0  # NA는 0으로 변환
}

idx <- match(z$납세자통합관리번호, .p$납세자통합관리번호)
for (v in setdiff(names(.p), '납세자통합관리번호'))
    z[[v]] <- p[idx, v]
```

○ 데이터 ⑤의 재무제표를 추가하고 결과를 저장함 : 각 연도의 여러 가지 비율과 원자료를 ①번 자료의 납세자통합관리번호와 재무제표상의 납세자통합관리번호를 이용하여 매칭시킴

```
w <- readRDS('재무제표.rds')
names(w)[1] <- '연도'

DEN <- function(x, eps = 1e-3) sign(x)*pmax(abs(x),eps)

r <- w[, 1:3]

r$부채비율 <- w[[103]]/DEN(w[[106]])
r$유동부채비율 <- w[[76]]/DEN(w[[106]])
r$차입금의존도 <- (w[[79]]+w[[84]]+w[[88]])/DEN(w[[75]])
r$단기성차입금비율 <- (w[[79]]+w[[84]])/
    DEN(w[[79]]+w[[84]]+w[[88]])
r$재고자산유동자산비율 <- w[[26]]/DEN(w[[6]])
r$당좌비율 <- w[[7]]/DEN(w[[76]])
r$현금성자산비율 <- (w[[8]]+w[[13]]+w[[14]]+w[[43]]+w[[44]])/
```

```
    DEN(w[[79]]+w[[84]])
r$영업이익률 <- w[[176]]/DEN(w[[108]])
r$매출액순이익률 <- w[[223]]/DEN(w[[108]])
r$금융비용매출액비율 <- (w[[200]]-w[[180]])/DEN(w[[108]])
r$재고자산회전율 <- w[[108]]/DEN(w[[26]])
r$매출채권회전율 <- w[[108]]/DEN(w[[9]])
r$매입채무회전율 <- w[[108]]/DEN(w[[77]])
r$운전자금회전율 <- w[[108]]/
    DEN(w[[9]]+w[[26]]+w[[77]]+w[[19]]-w[[81]])

DealNaN<-function(x) ifelse(is.finite(x),x,0)

r$유동자산금액 <- DealNaN(w[[6]])
r$당좌자산금액 <- DealNaN(w[[7]])
r$매출채권금액 <- DealNaN(w[[9]])
r$재고자산금액 <- DealNaN(w[[26]])
r$자산총계금액 <- DealNaN(w[[75]])
r$유동부채금액 <- DealNaN(w[[76]])
r$매입채무금액 <- DealNaN(w[[77]])
r$단기차입금액 <- DealNaN(w[[79]])
r$유동성장기부채금액 <- DealNaN(w[[84]])
r$장기차입금액 <- DealNaN(w[[88]])
r$선수금액 <- DealNaN(w[[81]])
r$선급금액 <- DealNaN(w[[19]])
r$부채총계금액 <- DealNaN(w[[103]])
r$자본총계금액 <- DealNaN(w[[106]])
r$매출금액 <- DealNaN(w[[108]])
r$영업손익금액 <- DealNaN(w[[176]])
r$현금성자산금액 <- DealNaN(w[[8]]+w[[13]]+w[[14]]+w[[43]]+w[[44]])
r$당기순손익금액 <- DealNaN(w[[223]])
r$이자비용금액 <- DealNaN(w[[200]])
r$이자수익금액 <- DealNaN(w[[180]])

for (j in sort(unique(w$연도))) {
  r1 <- r[r$연도==j,]
  idx <- match(z$납세자통합관리번호, r1$납세자통합관리번호)
  for (v in names(r1)[3:ncol(r1)]) z[[v %+% j]] <- r1[idx, v]
}

saveRDS(z, 'FinalDataSet.rds')
```

○ 이상의 과정을 거쳐 분석에 필요한 모든 데이터를 `FinalDataSet.rds`에 저장하였음

□ 데이터베이스에 당초부터 잘못된 정보가 입력되어 있을 가능성도 있으므로, 중간의 여러 단계에서 요약통계량 등을 이용하여 데이터의 무오류성(integrity)를 점검하여야 하며, 이 과정에서 매우 긴 시간이 소요될 수 있음에 유의할 것

제 3 장

체납 결정요인 분석

□ 본 연구에서는 납부기한 내 완납이 이루어지지 않아 체납될 확률과 체납이 발생하는 경우 체납 기간을 추정하는 모형을 제시하고자 함

□ 모형 추정에 이용한 데이터는 납세액이 5,000,000원 이상인 납세의무자들의 2015년 ~ 2017년 데이터임

□ 국세청 기초자료로부터 분석용 데이터를 생성하였으며, 데이터 생성 방법에 대해서는 2장 참조

3.1 모형과 방법론

□ (12개 모형) 탐색을 위하여 총 12개의 모형을 설정하고, 전체 데이터의 85%를 이용하여 각각의 모형을 추정함(나머지 15%는 12개 모형의 비교를 위한 검증용으로 사용함)

○ 크게 구분하여 2단(two-tier) 모형과 단일(one-tier) 모형 두 가지 종류의 모형을 사용

› 2단(two-tier) 모형은 체납 여부를 설명하는 '체납확률 추정 모형'과 체납 시 체납 기간을 설명하는 '체납기간 추정 모형'을 별도로 추정함

› 단일(one-tier) 모형은 완납까지 걸리는 시간을 설명하는 모형을 추정하고, 그 추정 결과를 바탕으로 체납 확률을 결정함

○ 2단 모형에서는 체납확률 추정 모형에서 체납 여부를 설명하기 위해 표준적인 '로짓'과 '프로빗' 두 모형을 고려하고, 체납기간 추정 모형에서 표준적인 Weibull 모형과 로그정규분포 모형을 고려하므로, 총 4가지 모형을 고려함(각 모형의 의미에 대해서는 3.4절 참조)

○ 단일 모형은 완납까지 걸리는 시간 모형으로서 표준적인 Weibull 모형과 로그정규분포 모형을 고려함(2가지)

○ 2단 모형 4개와 단일 모형 2개 각각에 대하여, 재무제표상의 설명변수로서 각종 비율을 사용하는 모형(비율형)과 재무제표상의 변수 자체에 로그를 취한 값을 사용하는 모형(로그형)을 고려함(이유에 대해서는 이하 3.4절 참조)

○ (2단 모형 4개 + 단일 모형 2개) × (비율형 또는 로그형) = 12개 모형

〈표 3.1〉 12개 모형 비교

모형	2단/단일 여부	모형 상세 내역	재무제표 변수
M1A	2단	로짓 + Weibull	비율형
M1B	2단	로짓 + Weibull	로그형
M2A	2단	로짓 + 로그정규분포	비율형
M2B	2단	로짓 + 로그정규분포	로그형
M3A	2단	프로빗 + Weibull	비율형
M3B	2단	프로빗 + Weibull	로그형
M4A	2단	프로빗 + 로그정규분포	비율형
M4B	2단	프로빗 + 로그정규분포	로그형
M5A	단일	Weibull	비율형
M5B	단일	Weibull	로그형
M6A	단일	로그정규분포	비율형
M6B	단일	로그정규분포	로그형

주. '2단' 모형은 체납여부를 설명하는 모형과 체납시 정리기간을 설명하는 모형을 별도로 가지고 있음. '단일' 모형은 납부기간을 설명하는 단일 모형이며 이로부터 체납과 체납정리기간을 도출함.

□ (피설명변수) 모형에 따라 다음과 같은 변수를 이용

○ 2단계 모형 : 체납여부(2단 모형의 '체납확률 추정 모형') 및 납기부터 완납까지 걸리는 시간(일, 2단 모형의 '체납기간 추정 모형'; 체납된 데이터만으로 한정하여 분석)

○ 단일 모형 : 고지된 시점부터 완납까지 걸리는 시간(일)

□ (설명변수) 설명변수로 시간추세, 차감고지세액(로그), 납세의무자의 특성 변수, 납세의무자의 납세 이력 관련 변수, 재무제표 사용

○ 시간추세는 시간에 따른 체납 성향의 변화를 포착하며, 차감고지세액도 체납 행태의 중요한 결정요인일 수 있으므로 포함시킴

○ 납세의무자의 특성으로 연령, 세대원수, 자동차보유여부, 명의위장사업자여부, 전년도 종합소득(로그)을 고려함

○ 납세의무자의 납세이력으로서, 전년도 유예신청건수, 직전 2개 연도의 평균 유예신청건수, 직전 3개 연도의 평균 유예신청건수, 전년도 고지서 개수를 사용함

○ 다양한 재무제표상 변수를 비율형과 로그형으로 각각 고려함

○ 상세한 내용은 3.4절 참조

□ (체납발생 여부 예측법) 총 12개 모형 각각에 대하여 각 관측치별 체납 발생 확률을 구하고, 이 확률이 어떤 문턱값보다 높으면 체납이 발생할 것으로 예측하고 문턱값보다 낮으면 체납이 발생하지 않을 것으로 예측함

□ (문턱값의 결정) 추정에 사용하지 않은 나머지 15% 데이터를 활용하여 각 문턱값별로 체납 여부들을 0과 1로 예측하고 나서, '체납된 고지가 모형에 의해 비체납으로 잘못 예측되는 비율(FN, false negative rate)'과 '비체납 고지가 모형에 의해 체납으로 잘못 예측되는 비율(FT, false positive rate)'을 각각 제곱하여 합한 값이 최소화되도록 문턱값들을 결정함

○ 문턱값을 높일 경우 체납이 발생하지 않을 것으로 예측하는 경향이 있으므로 '체납이 비체납으로 잘못 예측되는 비율(FN)'이 증가함

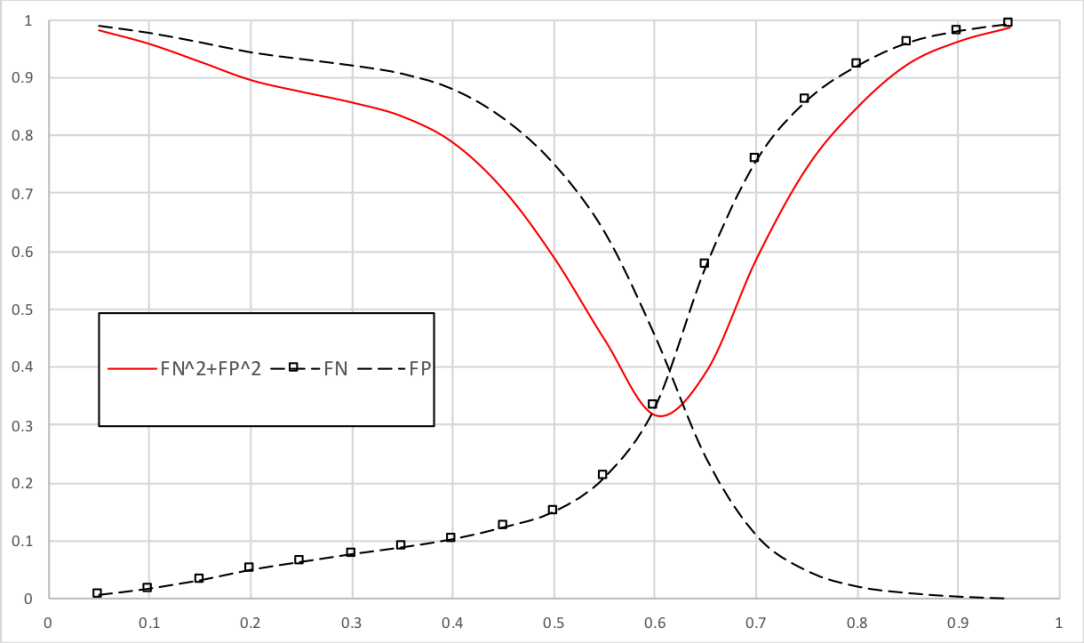
○ 문턱값을 낮출 경우 체납이 발생할 것으로 예측하기 쉬워지므로 '비체납이 체납으로 잘못 예측되는 비율(FP)'이 증가함

○ 두 오류 비율은 상충관계에 있으므로, 모두 감소시킬 수는 없음

○ FN과 FP의 제곱합을 최소화시키는 문턱값을 사용함

○ 예를 들어, 〈그림 3.1〉은 문턱값(횡축)에 따른 M6B 모형의 FP, FN, 그리고 $FN^2 + FP^2$의 패턴을 보여주고 있음. 두 비율의 제곱합은 문턱값이 약 0.6일 때 최소화됨

〈그림 3.1〉 문턱값에 따른 체납자와 비체납자의 올바른 예측 확률(M6B 모형)



□ (모형의 예측 성과 판단) 위의 방법으로 정한 문턱값을 사용하여 각각의 모형에서 체납 여부에 대한 예측성과와 체납정리기간에 대한 예측성과를 각 모형별로 측정함

○ 전체 데이터를 추정용 데이터와 검증용 데이터로 (85:15 비율로) 구분하는 것을 임의로 m회 반복하여(m = 30) 체납여부 예측성과와 체납정리기간 예측성과를 모형별로 측정한 후 이를 평균하여 판단함

○ **체납여부 예측성과** : 검증용 데이터에 대하여 두 예측오류 비율(FN과 FP)를 계산하고, $FN^2 + FP^2$이 작을수록 예측의 성과가 좋은 것으로 간주함

› 두 예측오류 비율 제곱의 단순합 대신에 가중합을 이용할 수도 있음(예를 들어 전체에서 체납된 고지와 비체납된 고지의 비중으로 가중합)

› 본 연구에서 분석에 이용한 데이터(납세액 5백만 원 이상)는 약 73%가 체납이 되지 않은 기록이므로, 이러한 가중합을 이용할 경우 항상 체납하지 않는다는 예측을 하더라도 예측의 성과가 약 73%가 되어 표면적으로는 예측 성과가 좋은 것처럼 보인다는 문제가 있음

› 체납 데이터와 기한 내 완납 데이터의 비중 차이가 크고, 체납자들에 대한 체납 예측이 중요한 목적이므로 가중합 대신 두 오류비율 제곱의 단순합을 이용하였음

○ **체납기간 예측성과 :** 예측용 데이터 중 체납된 고지에 대하여 (실제 체납 기간 – 예측된 체납 기간)의 제곱의 평균에 제곱근을 취한 평균제곱오차제곱근(root mean squared errors, RMSE)을 이용하여 측정함

› RMSE는 예측 연구에서 가장 널리 사용되는 지표임

› RMSE의 값이 작을수록 예측의 성과가 좋은 것임

□ (모형선정 및 추정 방법 요약) 이상에 설명한 모형 선정 방법을 순서대로 요약 설명하면 다음과 같음

1. 2장에 생성한 자료로부터 실제 분석에 필요한 구체적인 변수들을 생성함(이 경우 아래 3.2절의 설명과 같이 변수들을 적절하게 정규화함)
2. ⟨표 3.1⟩의 각 모형을 추정함
   ○ 2단 모형의 경우 체납확률 추정 모형을 '로짓'이나 '프로빗' 중 해당 방법을 이용하여 추정하고, 다음으로, 체납된 데이터만 활용하여 '체납기간 추정 모형'에 해당하는 Duration 모형을 추정함
   ○ 단일 모형의 경우 전체 데이터를 활용하여 해당 Duration 모형을 추정함
3. 앞의 제2단계에서 추정한 결과를 바탕으로 각 모형별로 각 관측치에 대하여 체납확률을 이하 ⟨표 3.5⟩의 $\Pr[Y = 1|x, \tau, T]$ 열 산식에 의하여 계산함
4. 각각의 모형에서, 앞의 제3단계에서 추정한 체납확률로부터 가능한 모든 문턱값에서 FN과 FP의 제곱합을 계산하고, 이 값을 최소화시키는 최적의 문턱값을 찾음
   ○ FN은 체납된 고지 중 비체납으로 잘못 예측된 고지의 비율, FP는 비체납된 고지 중 체납으로 잘못 예측된 비율을 의미함

○ 각 문턱값별로 FN은 체납된 고지 중 제3단계에서 구한 체납확률이 문턱값보다 작은 비율을 의미하며, FP는 비체납된 고지 중 제3단계에서 구한 체납확률이 문턱값보다 큰 비율을 의미함

5. 각각의 모형에서 앞의 제4단계에서 구한 최적의 문턱값을 이용하여 구한 오류비율 제곱합의 값을 12개 모형별로 비교함. 그 결과가 아래 〈표 3.2〉의 '체납여부 예측오차'에 해당함

6. 앞의 제2단계에서 추정한 결과를 바탕으로 각 모형별로 체납된 고지에 대하여 〈표 3.5〉의 마지막 열 산식에 따라 체납 시 체납기간 예측값을 구함

7. 각 모형별로, 앞의 제6단계에서 구한 체납기간 예측값과 실제 체납기간의 차이를 제곱한 값의 평균(체납된 고지에 한함)의 제곱근을 구하여 비교함. 그 결과가 아래 〈표 3.2〉의 '체납기간 예측오차'에 해당함

8. 앞의 제5단계와 제7단계에서 구한 체납여부 예측오차와 체납기간 예측오차를 바탕으로 목적에 맞추어 최적의 모형을 선정함

   ○ 만약 체납여부의 예측이 목적이라면 제5단계의 지표를 최소화시키는 모형을 선정

   ○ 만약 체납이 확정된 상태에서 체납기간의 예측이 목적이라면 제7단계의 지표를 최소화시키는 모형을 선정

   ○ 만약 체납여부와 체납 정리기간을 모두 어느 정도로 잘 예측하는 것이 목적이라면 이 두 지표를 종합하여 판단함

9. 선정된 최적 모형에 대하여 전체 데이터를 이용하여 추정함(추정방법은 앞의 제2단계에 설명된 바와 동일함)

□ (실제 예측) 본 연구의 방법을 이용하여 신규 납세고지 발생 시 체납여부를 예상하거나 체납된 고지건에 대하여 정리기간을 예측할 수 있음

○ 체납여부의 경우, 최종적으로 추정한 계수(앞의 제9단계 참조)와 해당 납세고지건의 특성들을 이용하여 체납확률을 구하고(해당 모형에 대하여 앞의 제3단계 방법 이용), 이 체납확률이 문턱값(앞의 제4단계 참조)보다 높으면 체납할 것으로 예측함

○ 체납된 고지건의 경우 최종적으로 추정한 계수(앞의 제9단계 참조)와 해당 납세고지건의 특성을 이용하여 체납기간을 예측함(해당 모형에 대하여 앞의 제6단계 방법 이용)

□ 본 연구의 목적은 본 연구를 위하여 국세청이 제공한 기초자료를 바탕으로 분석용 데이터를 생성하고 실제 분석을 진행하는 것이었으며, 본 연구의 내용을 실무에 적용하게 위해서는 더 광범위한 데이터와 모형을 이용하여 예측 목적에 맞는 분석을 선행할 것이 요구됨

○ 본 연구의 경우 국세청 제공 자료를 바탕으로 모형을 추정하여 요인별로 체납확률과 체납시 정리기간에 미치는 영향을 파악하는 것이 주 목적이었음

○ 실제 활용을 위해서는 더욱 방대한 데이터를 다양한 머신러닝 기법(신경망, 서포트 벡터 머신, 트리 등 여러 머신러닝 기법과 다양한 앙상블 기법)으로 분석하여 예측에 최적화된 모형을 탐색하여야 할 것임

○ 이 경우에도 본 연구의 전반적인 방법론을 따를 수 있을 것임

3.2 데이터 및 요약통계량

□ 데이터의 출처와 생성 방법에 대해서는 2장 참조

□ 요약통계량: 변수들의 평균을 분석에 사용한 전체 데이터, 체납으로 확인된 데이터, 체납이 아닌 것으로 확인된 데이터 등으로 구분해 변수들의 평균을 구하였음(부록 참조)

□ 추정시 변수들의 값의 차이가 너무 많이 날 경우 이론상으로는 문제가 없으나 현실적으로는 컴퓨터의 계산 한계의 문제로 인해 추정이 불가능해질 수 있음. 따라서 추정에는 $(\text{해당변수} - a)/b$의 형태로 적절한 정규화 조치를 취한 후 사용하였음(변수별로 사용한 a와 b에 대해서는 부록 참조)

3.3 추정 결과

□ 12개의 모형 중 M6B가 가장 적절한 것으로 판단함

○ 12개 모형들의 체납여부 성과와 체납기간 예측 성과는 다음 표와 같음.

〈표 3.2〉 모형별 종합적 예측 성과

모형	M1A	M1B	M2A	M2B
체납여부 예측오차	0.2778 (0.3)	0.2106 (0.3)	0.2278 (0.3)	0.2106 (0.3)
체납기간 예측오차	233.7537	253.5113	499.4739	503.7130
모형	M3A	M3B	M4A	M4B
체납여부 예측오차	0.2288 (0.2)	0.2097 (0.2)	0.2288 (0.2)	0.2097 (0.2)
체납기간 예측오차	233.7537	253.5113	499.4739	503.7130
모형	M5A	M5B	M6A	M6B
체납여부 예측오차	0.3096 (0.6)	0.2984 (0.6)	0.3109 (0.6)	0.3163 (0.6)
체납기간 예측오차	3842.4445	38381419.7618	156.3818	157.0806

출처 : 국세청 제공 데이터로부터 자체 계산

주 : '체납여부 예측오차'는 $FN^2 + FP^2$ 이며, '체납기간 예측오차'는 체납된 납세고지에 대하여 실제 체납기간과 예측값 간 차이 제곱의 평균의 제곱근(RMSE)을 의미함. 두 지표 모두 작을수록 좋은 모형임을 의미함. '체납여부 예측오차' 행에서 괄호 안의 값은 사용한 최적 문턱값.

○ 체납여부 예측성과 면에서는 로그형 재무제표 변수를 사용한 프로빗 모형인 M3B와 M4B의 오차비율 제곱합이 0.2097로 가장 우수함

○ 반면 체납 기간 예측 측면에서 M6A와 M6B보다 다른 모형들의 RMSE(작을수록 좋은 모형을 의미함)는 작게는 80 정도 더 크고(M3A의 경우) 그 다음으로 예측 오차가 작은 것도 100 정도 더 크며(M3B의 경우), 300 이상 차이가 나거나(M2A, M2B, M4A, M4B) 심지어 비교가 무의미한 경우도 있음(M5A, M5B). 따라서 **M6A와 M6B가 다른 모형에 비해 확연히 우수한 성과를 보여 주고 있음**

○ 비율형과 로그형을 비교하면, M6A가 M6B보다 체납 기간 예측 측면에서 근소하게 더 나으나 거의 무시할 만 하고, M6A에서 이용하는 비율형 재무제표 변수들에서 분모가 0이 되는 경우도 많았고 변수들의 값의 분산이 매우 커서 **로그형(M6B)을 이용하는 것이 더 적절한 것으로 판단됨**

○ 체납여부와 체납기간을 종합적으로 고려하면 M6B를 추정에 이용하는 모형으로 결

정하나, 체납여부의 예측만을 고려할 경우 M3B와 M4B에서 사용한 '프로빗' 모형도 좋음

□ M6B 모형의 추정 결과는 다음과 같음

〈표 3.3〉 M6B 모형의 추정 결과

설명변수	계수 추정값	t값	간단한 해석
상수항	3.11772	78.0979	
추세	0.08988	14.3979	체납확률 증가 추세
ln차감고지세액	0.09263	32.9708	클수록 체납확률 ↑
연령	-0.00343	-21.9541	젊을수록 체납확률 ↑
(세대원수)	(-0.00388)	(-1.8482)	
자동차보유여부	-0.16082	-25.7103	보유하면 체납확률 ↓
명의위장사업자여부	0.09304	1.9567	명의위장인 경우 체납확률 ↑
ln전년도종합소득	-0.23138	-33.2733	저소득일수록 체납확률 ↑
전년도유예신청건수	-0.09655	-3.7293	많을수록 체납확률 ↓
(직전2년평균유예신청건수)	(-0.10256)	(-1.8267)	많을수록 체납확률 ↓
직전3년평균유예신청건수	0.70183	12.1145	많을수록 체납확률 ↑
전년도고지서개수	0.06727	59.5218	많을수록 체납확률 ↑
ln전년도유동자산금액	0.71467	16.2250	클수록 체납확률 ↑
ln전년도당좌자산금액	-0.44184	-10.5822	클수록 체납확률 ↓
ln전년도매출채권금액	0.07343	16.7537	클수록 체납확률 ↑
ln전년도재고자산금액	-0.07148	-18.0844	클수록 체납확률 ↓
ln전년도자산총계금액	-0.32634	-12.2987	클수록 체납확률 ↓
ln전년도유동부채금액	0.09369	7.2743	클수록 체납확률 ↑
(ln전년도매입채무금액)	(0.00210)	(0.4833)	
ln전년도단기차입금액	0.01901	4.8427	클수록 체납확률 ↑

(다음 페이지에 계속)

설명변수	계수 추정값	t값	간단한 해석
ln전년도유동성장기부채금액	0.00593	2.1666	클수록 체납확률 ↑
ln전년도장기차입금액	0.04330	9.1073	클수록 체납확률 ↑
(ln전년도선수금액)	(-0.00443)	(-1.5317)	
ln전년도선급금액	-0.01042	-3.5011	클수록 체납확률 ↓
ln전년도부채총계금액	-0.14044	-8.8263	클수록 체납확률 ↓
ln전년도자본총계금액	-0.15820	-18.2015	클수록 체납확률 ↓
ln전년도매출금액	0.41520	17.9463	클수록 체납확률 ↑
(ln전년도영업이익금액)	(0.00306)	(0.0508)	
(ln전년도영업손실금액)	(-0.00393)	(-0.2269)	
ln전년도현금성자산금액	-0.03562	-11.7405	클수록 체납확률 ↓
ln전년도당기순이익금액	-0.14314	-2.5036	클수록 체납확률 ↓
(ln전년도당기순손실금액)	(-0.03471)	(-1.9596)	
ln전년도이자비용금액	-0.04645	-8.2313	클수록 체납확률 ↓
ln전년도이자수익금액	-0.04427	-15.0212	클수록 체납확률 ↓
$\ln(\sigma)$	-0.03262	-16.0673	$\hat{\sigma} = 0.968$
표본 크기	126,086		
모형의 우도함수값	-603,748.7		
우도함수값(상수항만 포함)	-609,312.4		

주. 괄호 안의 변수들은 5% 수준에서 유의하지 않으며, 나머지는 모두 5% 수준에서 유의함

○ 계수추정값이 양(+)인 변수의 경우 다른 조건이 동일한 상황에서 이 변수의 값이 커질수록 체납 확률과 체납정리기간이 증대됨을 의미. 반대로 계수추정값이 음(−)인 변수의 경우 다른 조건들이 동일한 상황에서 이 변수의 값이 커질수록 체납 확률과 체납정리기간이 감소됨을 의미

○ '추세'의 계수가 양(+)이고 통계적으로 유의함

› 다른 조건을 통제하고 나면 체납 확률이 점차 높아지고 체납정리기간이 증대되는 추세가 관측됨

○ ‘ln차감고지세액’의 계수추정값이 양(+)이고 통계적으로 유의적임

› 다른 조건이 동일하고 차감고지세액만 차이가 나는 두 납세의무자가 있을 경우 ‘차감고지세액’이 클수록 체납 확률이 높고 체납정리기간이 더 긺을 의미

○ ‘연령’의 계수추정값이 음(−)이고 통계적으로 유의적임

› 다른 조건이 동일하고 연령이 다른 두 납세의무자가 있을 경우 젊은 납세의무자의 체납 확률이 높고 체납정리기간이 더 긺을 의미

○ ‘ln전년도종합소득’의 계수추정값이 음(−)이고 통계적으로 유의적임

› 다른 조건이 동일하고 전년도 종합소득이 서로 다른 두 납세의무자가 있을 경우 전년도 종합소득이 작은 납세의무자의 체납 확률이 높고 체납정리기간이 더 긺을 의미

○ ‘과거 유예 신청 건수’의 경우 다음과 같이 해석됨

› ‘전년도유예신청건수’와 ‘직전2년유예신청건수’의 계수는 음(−)의 부호를 갖지만 ‘직전3년평균유예신청건수’의 계수값의 크기가 매우 크고 부호가 양(+)임. 전반적으로, 다른 조건이 동일한 경우, 과거 유예 신청 건수가 많을수록 체납 확률이 높고 체납정리기간이 더 긺을 의미

› ‘전년도유예신청건수’의 계수 부호가 음(−)인 것은 ‘직전2년평균유예신청건수’와 ‘직전3년평균유예신청건수’를 포함해 다른 모든 변수들의 값이 동일한 두 납세의무자가 전년도 유예 신청 건수만 다를 경우 전년도 유예 신청 건수가 많았던 사람이 체납 확률이 낮고 체납정리기간이 더 짧음을 의미함

· ‘직전2년평균유예신청건수’와 ‘직전3년평균유예신청건수’가 동일한 두 납세의무자 A와 B에 대해, A의 전년도 유예신청건수가 B의 유예신청건수보다 많음에도 불구하고 두 납세의무자의 직전 2년, 직전 3년 평균 유예신청 건수가 동일하다는 것은 A의 2년 전과 3년 전의 유예신청건수의 합이 B의 2년 전과 3년 전 유예신청 건수의 합보다 작다는 것을 의미함

· 따라서 이 추정결과에서 '전년도유예신청건수'의 계수 부호가 음(−)인 것은 과거 2년 또는 3년의 유예신청 건수가 작은 사람이 체납 확률이 낮고 체납정리기간이 더 짧음을 의미함

○ 다른 조건이 동일한 납세의무자 중 '전년도 고지서 개수'가 많은 쪽이 체납 확률이 높고 체납정리기간이 더 김

○ 재무제표 관련 변수들의 계수 추정값들은 '다른 변수의 값이 동일할 때' 해당 변수의 증감에 따라 체납 확률과 체납 정리기간이 어떻게 달라지는지를 알려 주는 것임. 경우에 따라서는 특정 변수의 계수 추정값을 해석할 때 재무제표 상 해당 변수들이 상호 연관성 등 여러 요소들을 고려하여야 할 수도 있음.

› 예컨대, 자산총계를 포함해 다른 모든 변수 값이 동일하고 전년도 당좌자산금액만 다른 두 납세의무자는 있을 수 없음.

› 뿐만 아니라 모형에 포함된 변수들의 값은 해당 변수의 실제 log값이 아니라 log 값을 정규화한 값이므로 계수 추정값을 정량적으로 해석하는 데에 추가적인 주의가 필요함

› 재무제표의 제반 항목들이 체납 확률과 체납 정리기간에 미치는 효과, 특히 정량적 효과는 계수추정값만으로 판단하여서는 안되고 필요한 값을 이용해 실제로 계산을 해 보아야 함

3.4 모형과 추정법의 수학적 설명

□ 본 절에서는 모형과 추정법을 수학적으로 설명함

□ 내용이 매우 기술적(technical)일 수 있음에 유의할 것

3.4.1 변수의 설명과 표기

□ 관측 단위는 징수결정ID로 필요 시 j 첨자로 표시함(의미가 분명한 경우에는 변수들에서 j 첨자를 생략하고 표기함)

□ τ: 해당 고지의 발행 일자

□ T: 최종납부기한

□ t: 해당 고지의 상태가 최종적으로 관측된 시점(본 연구에 사용한 데이터의 경우 모두 2018년 10월 9일)

□ S: 완납 시점. 데이터 관측 시점(t) 현재 완납되지 않은 고지인 경우 S는 알 수 없음

□ (D, c): D는 고지가 완납까지 걸리는 기간이며 정의상 $D = S - \tau$임. 다만, t시점 현재 완납되지 않은 고지인 경우 D가 관측되지 않음. c는 D의 관측여부를 나타내는 변수. 다음과 같이 정함

  ○ 완납된 시점 S가 관찰된 경우 (완료된 데이터): $c = 1, D = S - \tau + 1$

  ○ 관측 시점 t까지 완납되지 않은 경우(오른쪽 절단된 데이터): $c = 0$이며 이 경우 $D > t - \tau + 1$이지만 D값은 알 수 없음

□ y: 체납 여부. 체납된 경우 1, 기한 내 완납된 경우 0. 즉

$$y = I(D > T - \tau + 1).$$

여기서 $I(A)$는 지시함수로서 A가 참일 경우 1, 거짓일 경우 0임.

□ D^C: 해당 고지가 체납된 경우 완납까지 걸리는 시간(일)을 나타내는 변수. 체납되지 않은 경우 $D^C = 0$으로 정의. 따라서 $D^C = Dy$.

□ x: 납세고지시점 직전까지 가용한 설명변수들의 벡터. 모형 추정시 오류 발생을 최소화하기 위해 '추세' 변수를 제외한 여타 변수는 어떤 a와 b에 대해 $\frac{\text{변수}-a}{b}$와 같이 정규화(normalization)하여 추정에 이용함(정규화에 사용한 변수별 a와 b값은 부록 참조)

  ○ 추세 : 납세고지 연도 − 2010. 다른 조건을 통제한 이후에 사람들의 납세 패턴에 시간에 걸쳐 특정한 추세가 존재하는지 알 수 있음

  ○ ln 차감고지세액: 차감고지세액의 로그값

  ○ 해당 납세의무자의 특성 변수

    › 연령 : 납세고지가 발생한 시점에 해당 납세의무자의 연령

› 세대원수: 납세고지가 발생한 시점에 해당 납세의무자의 세대원 수

› 자동차보유여부: 납세고지가 발생한 시점에 해당 납세의무자의 자동차 보유 여부. 보유시 1, 미보유시 0

› 명의위장사업자여부: 명의위장 사업자인 경우 1, 그렇지 않은 경우 0

› ln 전년도종합소득: 납세고지 발생 시점 직전 해(calendar)의 ln전년도종합소득값

○ 해당 납세의무자의 납세 이력 관련 변수(이하에서 '직전 해', '직전 2개년' 등은 calendar year를 말함)

› 전년도유예신청건수: 납세고지 발생 시점 직전 해에 징수유예를 신청한 건수

› 직전2개년도평균유예신청건수: 납세고지 발생 시점 직전 2개년 연평균 징수유예 신청 건수

› 직전3개년도평균유예신청건수: 납세고지 발생 시점 직전 3개년 연평균 징수유예 신청 건수

› 전년도고지서개수: 납세고지 발생 시점 직전 해 동안 해당 납세자에게 고지된 징수결정ID의 개수

○ 재무제표 변수(비율형): 재무제표 데이터 중 분모가 없는 경우 0.001을 대입하여 계산하였고, 분자가 없는 경우 0으로 처리하였음(이하에서 '직전 해'는 calendar year를 말하며, 모두 납세고지가 발행된 납세자에 관한 변수임)

› 전년도부채비율: 납세고지 발생 시점 직전 해의 부채비율

› 전년도유동부채비율: 납세고지 발생 시점 직전 해의 유동부채 비율

› 전년도차입금의존도: 납세고지 발생 시점 직전 해의 차입금 의존도

› 전년도단기성차입금비율: 납세고지 발생 시점 직전 해의 단기성 차입금 비율

› 전년도재고자산유동자산비율: 납세고지 발생 시점 직전 해의 재고자산유동자산비율

› 전년도당좌비율: 납세고지 발생 시점 직전 해의 당좌비율

› 전년도현금성자산비율: 납세고지 발생 시점 직전 해의 현금성자산비율

› 전년도영업이익률: 납세고지 발생 시점 직전 해의 영업이익률

› 전년도매출액순이익률 : 납세고지 발생 시점 직전 해의 매출액순이익률

› 전년도금융비용매출액비율 : 납세고지 발생 시점 직전 해의 금융비용매출액비율

› 전년도재고자산회전율 : 납세고지 발생 시점 직전 해의 재고자산회전율

› 전년도매출채권회전율 : 납세고지 발생 시점 직전 해의 매출채권회전율

› 전년도매입채무회전율 : 납세고지 발생 시점 직전 해의 매입채무회전율

› 전년도운전자금회전율 : 납세고지 발생 시점 직전 해의 운전자금회전율

○ 재무제표 변수(로그형): 재무제표상의 변수의 로그값(변수의 값이 0인 경우 0으로 설정)

› ln전년도유동자산금액

› ln전년도당좌자산금액

› ln전년도매출채권금액

› ln전년도재고자산금액

› ln전년도자산총계금액

› ln전년도유동부채금액

› ln전년도매입채무금액

› ln전년도단기차입금액

› ln전년도유동성장기부채금액

› ln전년도장기차입금액

› ln전년도선수금액

› ln전년도선급금액

› ln전년도부채총계금액

› ln전년도자본총계금액

› ln전년도매출금액

› ln전년도영업이익금액

› ln전년도영업손실금액

› ln전년도현금성자산금액

› ln전년도당기순이익금액

› ln전년도당기순손실금액

› ln전년도이자비용금액

› ln전년도이자수익금액

□ 설명변수 요약

○ 비율형 : 추세+개인 특성 변수+납세 이력 관련 변수+재무제표 변수(비율형)

○ 로그형 : 추세+개인 특성 변수+납세 이력 관련 변수+재무제표 변수(로그형)

3.4.2 2단 모형(Two-Tier Model)

□ 2단 모형은 체납확률을 추정하는 모형('체납확률 추정모형')과 체납이 이루어진 자료를 바탕으로 체납기간을 추정하는 모형('체납기간 추정모형')을 각각 별도로 추정하는 모형

□ **체납확률 추정모형**

○ 공변량 x가 주어졌을 때 체납 여부(y)의 조건부 확률 $\Pr[y|x]$를 다음과 같이 정의

$$\Pr[y=1|x] = \begin{cases} \Phi(x'\gamma) & \text{probit 모형의 경우} \\ \Lambda(x'\gamma) & \text{logit 모형의 경우} \end{cases} \tag{3.1}$$

이때 $\Phi(\cdot)$는 표준정규분포의 누적확률분포함수로서 $\Phi(s) = \int_{-\infty}^{s} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz$이며, $\Lambda(\cdot)$는 표준로지스틱분포의 누적확률분포함수로서 $\Lambda(s) = e^s/(1+e^s)$. $x'\gamma$는 x가 $x_1, \ldots, x_k$의 벡터일 때 $\gamma_0 + \gamma_1 x_1 + \cdots + \gamma_k x_k$를 뜻함.

○ 다음의 우도함수(likelihood function)을 극대화하는 최우추정법(maximum likelihood estimation, MLE)을 통해 γ의 추정량 $\hat{\gamma}$를 구함

$$L_B(\gamma) = \sum_j y_j \ln \Pr[y_j = 1|x_j] + (1-y_j)\ln(1 - \Pr[y_j = 1|x_j]) \tag{3.2}$$

○ $\hat{\gamma}$을 구한 후 이를 이용해 공변량의 값이 x인 고지의 체납 확률의 추정량($\widehat{\Pr[y=1|x]}$)은 다음과 같이 구함

$$\widehat{\Pr[y=1|x]} = \begin{cases} \Phi(x'\hat{\gamma}) & \text{probit 모형의 경우} \\ \Lambda(x'\hat{\gamma}) & \text{logit 모형의 경우} \end{cases} \tag{3.3}$$

○ 체납 여부 추정 : 공변량 값이 x인 고지가 체납될 것인지 여부($y(x)$)는 $\widehat{\Pr[y=1|x]}$이 어떤 문턱값(threshold value)보다 큰지 여부로 추정. 즉,

$$\hat{y}(x) = I(\widehat{\Pr[y=1|x]} > \text{정해진 문턱값}) \tag{3.4}$$

본 연구에서 사용한 문턱값의 결정 과정은 3.4.5절에서 설명함

□ **체납기간 추정모형**

○ 공변량 값이 x인 체납된 고지의 체납 기간 D^C가 0보다 큰 실수값을 갖는 연속형 확률변수이고 조건부 확률분포 $\Pr[D^C \leq s|x] = F(s|x;\theta)$(관련된 확률밀도함수는 $f(s|x;\theta)$)를 따른다고 가정

○ 체납된 데이터만 이용해 추정하므로 $S > T$임

○ 2018년 10월까지 완납 시점이 관측된($c=1$) 경우 $D^C = S - T$임. 반면 2018년 10월까지 완납이 되지 않은($c=0$) 관측값의 경우 $D^C > t - T$ (여기서 t는 데이터 추출 시점). 따라서 다음과 같이 절단된 표본의 로그우도함수를 구성할 수 있음

$$\begin{aligned} L_F(\theta) = \sum_{j:y_j=1} & I(c_j=0)\ln f(S_j - T_j|x_j;\theta) \\ & + I(c_j=1)\ln\left[1 - F(t_j - T_j|x_j;\theta)\right] \end{aligned} \tag{3.5}$$

○ 모수의 MLE추정량 $\hat{\theta}$가 구해지면 발행일자가 τ이고 납부기한이 T이며 공변량이 x인 고지가 체납되는 경우 평균 체납기간 $E[D^C|x]$의 추정값은

$$\widehat{E[D^C|x]} = \int_0^\infty s f(s|x;\hat{\theta})ds \tag{3.6}$$

와 같이 구할 수 있음.

○ 한편, 고지일자가 τ이고 납부기한이 T이며 공변량이 x인 고지가 체납되는 경우 완납시점의 추정값 $\hat{S}(x)$는

$$\hat{S}(x) = T + \lceil \widehat{E(D^C|x)} \rceil \tag{3.7}$$

여기서 $\lceil A \rceil$은 A보다 크거나 같은 정수 중 최솟값임

○ Hazard 분석

› Hazard 함수는 τ 시점에 발생된 최종납부기한 T의 고지가 $T+d$일까지 체납이 지속되다 $T+d+1$일에 완납할 확률과 관계된 함수임

› **체납기간 추정모형**과 관련된 hazard함수를 $\lambda_F(\cdot)$로 나타내면,

$$\lambda_F(d|x) = \lambda_F(d|x;\beta) = \frac{\lim_{\Delta\to 0} \frac{\Pr[d<D^C\le d+\Delta|x]}{\Delta}}{\Pr[D^C > d|x]} = \frac{f(d|x;\theta)}{1-F(d|x;\theta)} \tag{3.8}$$

로 정의되며 이 함수의 추정량으로

$$\hat{\lambda}_F(d|x) = \lambda_F(d|x;\hat{\beta}) = \frac{f(d|x;\hat{\theta})}{1-F(d|x;\hat{\theta})} \tag{3.9}$$

을 이용할 수 있음

› 체납기간이 d인 고지들 중 $d+\Delta$일 이내에 납부를 완료할 것으로 예상되는 건들의 비율을 알고 싶을 때

$$\frac{\sum_j \hat{\lambda}_F(d|x_j) \times \Delta \times (T_j + d \text{ 일 현재 체납 여부})}{\sum_j (T_j + d \text{ 일 현재 체납 여부})} \tag{3.10}$$

와 같이 근사 계산할 수 있음 (단, 여기서 Δ는 매우 작은 값이여야 함. 예컨대 $\Delta = 3$(3일이내) 등.)

› 그 외에 필요한 경우 hazard 함수를 분석하여 체납기간이 길어질수록 그에 비례해 납부 확률이 증가하는지 또는 오히려 감소하는지 등을 파악할 수 있음

□ 체납 기간 추정 모형에 이용한 조건부 분포함수들 : $s > 0$에 대해

○ Weibull 분포 (scale parameter = $e^{x'\beta}$, shape parameter = $\frac{1}{\sigma}$)

$$F(s|x;\beta,\sigma) = 1 - \exp\left(-\left(\frac{s}{e^{x'\beta}}\right)^{\frac{1}{\sigma}}\right), \tag{3.11}$$

$$f(s|x;\beta,\sigma) = \frac{e^{-x'\beta}}{\sigma}\left(\frac{s}{e^{x'\beta}}\right)^{\frac{1}{\sigma}-1}\exp\left(-\left(\frac{s}{e^{x'\beta}}\right)^{\frac{1}{\sigma}}\right), \tag{3.12}$$

$$\lambda_F(s|x;\beta,\sigma) = \frac{1}{\sigma s}\left(\frac{s}{e^{x'\beta}}\right)^{\frac{1}{\sigma}}, \tag{3.13}$$

$$E\left[D^C|x\right] = e^{x'\beta}\Gamma\left(1+\sigma\right) \tag{3.14}$$

여기서 $\Gamma(a)$는 감마함수로 $\Gamma(a) = \int_0^\infty e^{-t}t^{a-1}dt$로 정의됨

○ Lognormal 분포 (평균 $x'\beta$, 분산 σ)

$$F(s|x;\beta,\sigma) = \Phi\left(\frac{\ln s - x'\beta}{\sigma}\right) = \int_{-\infty}^{\frac{\ln s - x'\beta}{\sigma}} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}s^2\right)ds, \tag{3.15}$$

$$f(s|x;\beta,\sigma) = \frac{1}{s\sigma}\phi\left(\frac{\ln s - x'\beta}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma s}\exp\left(-\frac{1}{2}\left(\frac{\ln s - x'\beta}{\sigma}\right)^2\right), \tag{3.16}$$

$$\lambda_F(s|x;\beta,\sigma) = \frac{\phi\left(\frac{\ln s - x'\beta}{\sigma}\right)}{s\sigma\Phi\left(\frac{\ln s - x'\beta}{\sigma}\right)}, \tag{3.17}$$

$$E\left[D^C|x\right] = \exp\left(x'\beta + \frac{1}{2}\sigma^2\right) \tag{3.18}$$

3.4.3 단일 모형(One-Tier Model)

□ 2단 모형에서는 체납 확률과 체납된 고지의 체납 기간을 각각 별개의 모형으로 추정하는 데 비해, **단일 모형**에서는 고지시점부터 완납까지 걸리는 기간(D)을 추정하고 이로부터 체납 확률을 계산

□ 완납까지 걸리는 기간 D가 0보다 큰 실수값을 갖는 연속형 확률변수이고 공변량 x가 주어졌을 때 조건부 확률분포 $\Pr[D \le s|x] = H(s|x;\theta)$를 따르고 그 확률밀도함수가 $h(s|x;\theta)$라고 가정

□ $c=1$인 관측값(완료된 관측값)의 경우 $D = S - \tau$이고 $c=0$인 관측값(오른쪽 절단된 관측값)의 경우 $D > t - \tau$임. 따라서 다음과 같이 절단된 표본의 로그우도함수를 구성할 수 있음

$$L_H(\theta) = \sum_j I(c_j = 1) \ln h(S_j - \tau_j | x_j; \theta) + I(c_j = 0) \ln [1 - H(t_j - \tau_j | x_j; \theta)] \quad (3.19)$$

○ 완납일자가 고지일자와 동일한($S = \tau$) 경우 $D = 0$이 되고, 납부까지 걸리는 시간을 연속형 변수로 모형화하고 있으므로 $D = 0$을 그대로 이용하는 것이 타당하지 않은 측면이 있음. 따라서 관측된 완납일자가 고지일자와 같은 경우 $S = \tau + \frac{1}{2}$로 간주함

□ 모수의 MLE추정량 $\hat{\theta}$가 구해지면 이를 다음과 같이 이용할 수 있음

○ 고지일자가 τ, 납부 기한이 T이고 공변량이 x인 고지가 기한 내 납부될 확률 추정 :

$$\widehat{\Pr[D \leq T - \tau | x]} = H(T - \tau | x; \hat{\theta}) \quad (3.20)$$

○ 고지일자가 τ, 납부 기한이 T이고 공변량이 x인 고지가 체납될 확률 추정 :

$$\widehat{\Pr[D > T - \tau | x]} = 1 - H(T - \tau | x; \hat{\theta}) \quad (3.21)$$

› 2단 모형에서 체납 확률이 τ나 T의 함수가 아닌 반면 단일 모형에서는 $T - \tau$의 (감소)함수가 됨.

○ 고지일자가 τ, 납부 기한이 T이고 공변량이 x인 고지의 평균 납부 기간 추정 :

$$\widehat{E[D|x]} = \int_0^\infty s\, h(s|x; \hat{\theta}) ds \quad (3.22)$$

○ 고지일자가 τ, 납부 기한이 T이고 공변량이 x인 고지의 납부 시점 추정 :

$$\hat{S}(x, \tau, T) = \left\lceil \widehat{E[D|x]} \right\rceil + \tau \quad (3.23)$$

○ 고지일자가 τ, 납부 기한이 T이고 공변량이 x인 고지의 체납 여부 추정 : $\Pr[D > T - \tau | x]$ $(= 1 - H(T - \tau | x; \theta))$의 추정값이 주어진 문턱값 이상일 경우 체납될 것으로 추정. 즉,

$$\hat{y}(x, \tau, T) = I(1 - H(T - \tau | x; \hat{\theta}) > \text{주어진 문턱값}). \quad (3.24)$$

본 연구에서 사용한 문턱값의 결정 과정은 3.4.5소절에서 설명함

○ 고지일자가 τ, 납부 기한이 T이고 공변량이 x인 고지가 체납될 경우 완납까지 걸리는 평균 기간

$$E[D|x, \widehat{D > T - \tau}] = \frac{1}{1 - H(T - \tau|x; \hat{\theta})} \int_{T-\tau}^{\infty} s\, h(s|x; \hat{\theta}) ds. \tag{3.25}$$

○ 고지일자가 τ, 납부 기한이 T이고 공변량이 x인 고지가 체납될 경우 납부 시점 추정

$$\hat{S}(x, \tau, T) = \left\lceil E[D|x, \widehat{D > T - \tau}] \right\rceil + T \tag{3.26}$$

□ Hazard 분석

○ 2단 모형에서와 유사하게 hazard 함수를 정의하고 분석에 이용할 수 있음

○ 단일 모형에서 hazard 함수는 τ일에 고지된, 납부 기한이 T인 세금을 $\tau + s$일까지 납부를 하지 않다가 $\tau + s + 1$일에 납부할 확률과 관계된 함수임. 이를 $\lambda_H(\cdot)$로 나타내면,

$$\lambda_H(s|x) = \lambda_H(s|x; \beta) = \frac{\lim_{\Delta \to 0} \frac{\Pr[s < D \leq s + \Delta | x]}{\Delta}}{\Pr[D > s|x]} = \frac{h(s|x; \theta)}{1 - H(s|x; \theta)} \tag{3.27}$$

로 정의되며 이 함수의 추정량으로

$$\hat{\lambda}_H(s|x) = \lambda_H(s|x; \hat{\beta}) = \frac{h(s|x; \hat{\theta})}{1 - H(s|x; \hat{\theta})} \tag{3.28}$$

을 이용할 수 있음

○ 고지일로부터 s까지 미납한 고지들 중 $s + \Delta$이내에 납부를 완료할 것으로 예상되는 건들의 비율을 알고 싶을 때

$$\frac{\sum_j \hat{\lambda}_H(s|x_j) \times \Delta \times I(\tau_j + s \text{ 현재 시점 미납})}{\sum_j I(\tau_j + s \text{ 현재 시점 미납})} \tag{3.29}$$

와 같이 근사 계산할 수 있음 (여기서도 Δ는 매우 작은 값이어야 함)

○ **2단 모형**의 **체납 기간 추정 모형**에서와 마찬가지로, 필요할 경우, 체납기간과 납부확률의 관계를 분석할 수도 있음

□ 분석에 이용한 조건부 분포함수들: $s > 0$에 대해 가장 널리 사용되는 Weibull 분포와 로그정규(lognormal)분포를 사용함

○ Weibull 분포 (scale parameter = $e^{x'\beta}$, shape parameter = $\frac{1}{\sigma}$)

$$H(s|x;\beta,\sigma) = 1 - \exp\left(-\left(\frac{s}{e^{x'\beta}}\right)^{\frac{1}{\sigma}}\right), \tag{3.30}$$

$$h(s|x;\beta,\sigma) = \frac{e^{-x'\beta}}{\sigma}\left(\frac{s}{e^{x'\beta}}\right)^{\frac{1}{\sigma}-1}\exp\left(-\left(\frac{s}{e^{x'\beta}}\right)^{\frac{1}{\sigma}}\right), \tag{3.31}$$

$$\lambda_H(s|x;\beta,\sigma) = \frac{1}{\sigma s}\left(\frac{s}{e^{x'\beta}}\right)^{\frac{1}{\sigma}}, \tag{3.32}$$

$$E\left[D|x\right] = e^{x'\beta}\Gamma\left(1+\sigma\right), \tag{3.33}$$

$$\Pr[y=1|x,\tau,T] = \exp\left(-\left(\frac{T-\tau}{e^{x'\beta}}\right)^{\frac{1}{\sigma}}\right), \tag{3.34}$$

$$\begin{aligned} E[D|x, D>T-\tau] &= \frac{\int_{T-\tau}^{\infty} s h(s|x;\beta,\sigma)ds}{1-H\left(T-\tau|x;\beta,\sigma\right)} \\ &= \frac{e^{x'\beta}}{1-H\left(T-\tau|x;\beta,\sigma\right)}\int_{\left(\frac{T-\tau}{e^{x'\beta}}\right)^{\frac{1}{\sigma}}}^{\infty} w^{\sigma}e^{-w}dw \\ &= \frac{E[D|x]}{1-H\left(T-\tau|x;\beta,\sigma\right)}\frac{\Gamma\left(1+\sigma,\left(\frac{T-\tau}{e^{x'\beta}}\right)^{\frac{1}{\sigma}}\right)}{\Gamma\left(1+\sigma\right)} \\ &= \exp\left(x'\beta+\left(\frac{T-\tau}{e^{x'\beta}}\right)^{\frac{1}{\sigma}}\right)\Gamma\left(1+\sigma,\left(\frac{T-\tau}{e^{x'\beta}}\right)^{\frac{1}{\sigma}}\right) \end{aligned} \tag{3.35}$$

여기서 $\Gamma(a,b)$는 불완전감마함수로 $b>0$에 대해 $\Gamma(a,b)=\int_b^{\infty}e^{-t}t^{a-1}dt$로 정의됨

› $D|x \sim$ Weibull$(e^{x'\beta},\frac{1}{\sigma})$은 $(\frac{D}{e^{x'\beta}})^{\frac{1}{\sigma}}|x \sim Exponential(1)$ 임을 의미

○ Lognormal 분포 (평균 $x'\beta$, 분산 σ)

$$H(s|x;\beta,\sigma) = \Phi\left(\frac{\ln s - x'\beta}{\sigma}\right), \tag{3.36}$$

$$h(s|x;\beta,\sigma) = \frac{1}{s\sigma}\phi\left(\frac{\ln s - x'\beta}{\sigma}\right), \tag{3.37}$$

$$\lambda_H(s|x;\beta,\sigma) = \frac{\phi\left(\frac{\ln s - x'\beta}{\sigma}\right)}{s\sigma\Phi\left(\frac{\ln s - x'\beta}{\sigma}\right)}, \tag{3.38}$$

$$E[D|x] = \exp\left(x'\beta + \frac{1}{2}\sigma^2\right), \tag{3.39}$$

$$\Pr[y=1|x,\tau,T] = 1 - H(T-\tau|x;\beta,\sigma) = 1 - \Phi\left(\frac{\ln(T-\tau) - x'\beta}{\sigma}\right), \tag{3.40}$$

$$E[D|x, D > T-\tau] = \frac{\int_{T-\tau}^{\infty} \frac{s\phi\left(\frac{\ln s - x'\beta}{\sigma}\right)}{s\sigma} ds}{1 - \Phi\left(\frac{\ln(T-\tau) - x'\beta}{\sigma}\right)} = \frac{e^{x'\beta}\int_{\frac{\ln(T-\tau)-x'\beta}{\sigma}}^{\infty} e^{\sigma z}\phi(z)dz}{1 - \Phi\left(\frac{\ln(T-\tau)-x'\beta}{\sigma}\right)}$$
$$= E[D|x]\frac{\Phi\left(\frac{x'\beta - \ln(T-\tau)}{\sigma} + \sigma\right)}{\Phi\left(\frac{x'\beta - \ln(T-\tau)}{\sigma}\right)} \tag{3.41}$$

› $D|x \sim Lognormal(x'\beta, \sigma^2)$은 $\ln D|x \sim N(x'\beta, \sigma^2)$임을 의미함

□ 모형의 구분

○ 확률분포에 대한 가정과 x의 유형에 따라 다음과 같은 12가지 모형을 구성할 수 있음

⟨표 3.4⟩ 모형의 구분

모형 식별 부호		체납/납부 기간 확률	
		Weibull	Lognormal
2단 모형	체납여부 확률= Logit	M1A, M1B	M2A, M2B
	체납여부 확률= Probit	M3A, M3B	M4A, M4B
단일 모형		M5A, M5B	M6A, M6B

주: A는 설명변수들이 비율형임을, B는 설명변수들이 로그형임을 말함

○ 총 12개 모형 중 예측력이 가장 좋은 모형을 실제 추정 모형으로 선정하고자 함

3.4.4 체납 확률, 체납 기간, 성실 납부 성향의 추정

□ 3.4.2절과 3.4.3절의 내용을 요약하면 (공변량, 고지일자, 최종납부기한)이 (x, τ, T)인 고지의 체납 확률(Pr$[y = 1|x, \tau, T]$)과 체납 시 체납기간의 기댓값($E[D^C|x, \tau, T, y = 1]$)은 ⟨표 3.4⟩에 나열된 모형별로 다음과 같음

⟨표 3.5⟩ 모형별 체납 확률과 체납될 때 평균 체납 기간

모형	Pr$[y = 1\|x, \tau, T]$	$E[D^C\|x, \tau, T, y = 1]$
M1A, M1B	$\Phi(x'\gamma)$	$\exp(x'\beta)\Gamma(1+\sigma)$
M2A, M2B	$\Phi(x'\gamma)$	$\exp(x'\beta + \frac{1}{2}\sigma^2)$
M3A, M3B	$\Lambda(x'\gamma)$	$\exp(x'\beta)\Gamma(1+\sigma)$
M4A, M4B	$\Lambda(x'\gamma)$	$\exp(x'\beta + \frac{1}{2}\sigma^2)$
M5A, M5B	$\exp(-\alpha)$	$\exp(x'\beta + \alpha)\Gamma(1+\sigma, \alpha)$
M6A, M6B	$1 - \Phi(\ln\alpha)$	$\exp(x'\beta + \frac{1}{2}\sigma^2)\,\Phi(-\ln\alpha + \sigma)/\Phi(-\ln\alpha)$

주. $\Phi(\cdot)$는 표준정규분포의 누적확률분포함수, $\Lambda(z) = \dfrac{e^z}{1+e^z}$, $\Gamma(z, b) = \displaystyle\int_b^\infty t^{z-1}e^{-t}dt$, $\Gamma(z) = \Gamma(z, 0)$, $\alpha = \left[\dfrac{T-\tau}{\exp(x'\beta)}\right]^{1/\sigma}$.

□ 납세 의무자의 성실 납부 성향

○ 납세 의무자의 성실 납부 성향 점수를 납세 의무자가 체납하지 않을 확률로 정의해 볼 수 있음

○ 공변량이 x인 한 납세자가 고지일자가 τ_ℓ이고 최종납부기한이 T_ℓ인 $\ell = 1, 2, \cdots, L$ 개의 세금을 납부해야 할 때 이 납세자가 하나도 체납하지 않을 확률은 다음과 같음

$$\prod_{\ell=1}^{L}(\text{고지 } \ell\text{이 체납되지 않을 확률})$$

› 이 식은 L개의 세금 납부가 서로 독립적이라는 가정 하에 도출되었음

○ ⟨표 3.4⟩의 개별 고지의 체납 확률과 체납시 체납 기간을 이용해 성실 납부 성향 점수를 구하면 다음과 같음

⟨표 3.6⟩ 모형별 납세 의무자의 성실 납부 성향 점수

모형	성실 납부 성향 점수
M1A, M1B	$\prod_{\ell=1}^{L}\left[1-\Phi(x_\ell'\gamma)\right]$
M2A, M2B	
M3A, M3B	$\prod_{\ell=1}^{L}\left[1-\Lambda(x_\ell'\gamma)\right]$
M4A, M4B	
M5A, M5B	$\prod_{\ell=1}^{L}\left[1-\exp(-\alpha_\ell)\right]$
M6A, M6B	$\prod_{j=1}^{L}\Phi(\ln\alpha_\ell)$

주. $\Phi(\cdot)$는 표준정규분포의 누적확률분포함수, $\Lambda(z)=e^z/(1+e^z)$, $\alpha_\ell=[(T_\ell-\tau_\ell)/\exp(x_\ell'\beta)]^{1/\sigma}$

› 이렇게 정의된 성실 납부 성향 점수는 0과 1 사이의 값을 가짐.

› 성실 납부 성향 점수가 1에 가까운 값을 가질수록 해당 납세 의무자가 체납할 확률이 낮음

› 성실 납부 성향 점수가 0에 가까운 값을 가질수록 해당 납세 의무자가 1건 이상 체납할 확률이 높음

3.4.5 모형 추정의 세부 내용: 문턱값과 추정 모형의 선정

□ ⟨표 3.4⟩에 나열한 총 12개의 모형 중 예측력이 좋은 모형을 선정

□ 모형의 선정 절차

○ 가용한 자료를 학습용 데이터(training data)와 검증용 데이터(validation data)로 분할

○ 학습용 데이터(training data)를 이용해 각각의 모형의 모수들(γ, β, σ)을 추정한 후 추정된 모수 $\hat{\gamma}$, $\hat{\beta}$, $\hat{\sigma}$를 이용해 검증용 데이터(validation set)에 있는 납세고지건들에 대해 체납확률과 체납 시 체납정리기간을 추정.

⟨표 3.7⟩ 체납확률과 체납 시 체납기간 기댓값의 추정

모형	체납확률 추정값	체납 시 체납기간 추정값
M1A, M1B	$\Phi(x'\hat{\gamma})$	$\exp(x'\hat{\beta})\Gamma(1+\hat{\sigma})$
M2A, M2B	$\Phi(x'\hat{\gamma})$	$\exp(x'\hat{\beta}+\frac{1}{2}\hat{\sigma}^2)$
M3A, M3B	$\Lambda(x'\hat{\gamma})$	$\exp(x'\hat{\beta})\Gamma(1+\hat{\sigma})$
M4A, M4B	$\Lambda(x'\hat{\gamma})$	$\exp(x'\hat{\beta}+\frac{1}{2}\hat{\sigma}^2)$
M5A, M5B	$\exp(-\hat{\alpha})$	$\exp(x'\hat{\beta}+\hat{\alpha})\Gamma(1+\hat{\sigma},\hat{\alpha})$
M6A, M6B	$1-\Phi(\ln\hat{\alpha})$	$\exp(x'\hat{\beta}+\frac{1}{2}\hat{\sigma}^2)\,\Phi(-\ln\hat{\alpha}+\hat{\sigma})/\Phi(-\ln\hat{\alpha})$

주. ⟨표 3.5⟩의 모수들을 추정값으로 변경하여 구함. $\Phi(\cdot)$는 표준정규분포의 누적확률분포함수, $\Lambda(z)=\dfrac{e^z}{1+e^z}$, $\Gamma(z,b)=\displaystyle\int_b^\infty t^{z-1}e^{-t}dt$, $\Gamma(z)=\Gamma(z,0)$, $\hat{\alpha}=\left[\dfrac{T-\tau}{\exp(x'\hat{\beta})}\right]^{1/\hat{\sigma}}$.

› ⟨표 3.7⟩에 의하여 체납확률을 추정한 후 고지 j의 체납 여부의 예측값 $\hat{y}_j$은 다음과 같이 정함

$$\hat{y}_j(\text{문턱값}) = I(\text{고지 } j\text{의 체납확률 예측값} > \text{문턱값}) \tag{3.42}$$

○ 검증용 데이터를 이용해 예측한 후, 예측의 성과가 좋은 모형을 최종 추정 모형으로 선정. 최종 추정 모형이 선정된 후 가용한 모든 자료를 이용해 모형의 모수를 재 추정

○ 검증용 데이터는 전체의 약 15%로 정하였으며, 체납여부가 관측되도록 제약을 가한 후 임의추출함

○ 문턱값은 검증용 데이터에 있는 관측값들에 대해 다음과 같은 척도를 구해 결정

$$\text{FN} = 1 - \frac{\sum_j y_j\hat{y}_j(\text{문턱값})}{\sum_j y_j} \tag{3.43}$$

$$\text{FP} = 1 - \frac{\sum_j (1-y_j)(1-\hat{y}_j(\text{문턱값}))}{\sum_j (1-y_j)} \tag{3.44}$$

› FN은 검증용 데이터 내 체납된 건들 중 체납되지 않은 것으로 잘못 예측된 건들의 비율(false negative rates)이고, FP는 검증용 데이터 내 체납되지 않은 건들 중 체납된 것으로 잘못 예측된 건들의 비율(false positive rates). 이 두 오류 비율이

작을수록 체납 또는 기한 내 완납의 경우 각각에 대해 예측이 정확하게 이루어지는 것을 의미함

› FN과 FP는 문턱값에 따라 달라지므로, 문턱값을 p라 할 때 FN(p)와 FP(p)로 표현할 수도 있음

○ 각각의 모형에서 이 두 오류 비율 제곱의 합인

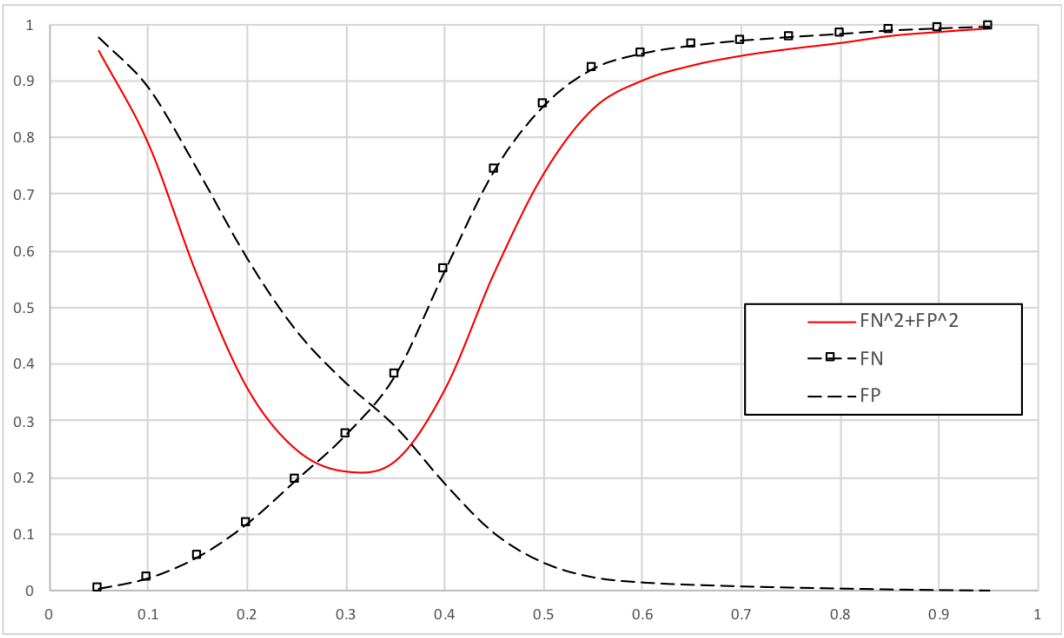
$$\text{FP}(p)^2 + \text{FN}(p)^2$$

을 최소화시키는 p값을 문턱값으로 선정

○ 문턱값이 커질수록 FN은 커지고 FP는 작아짐. 모형에 따라 FN과 FP의 구체적인 패턴은 달라짐(아래 그림 참조). 아래 그림들에서 횡축은 문턱값, 종축은 FN, FP, $\text{FN}^2 + \text{FP}^2$ 임

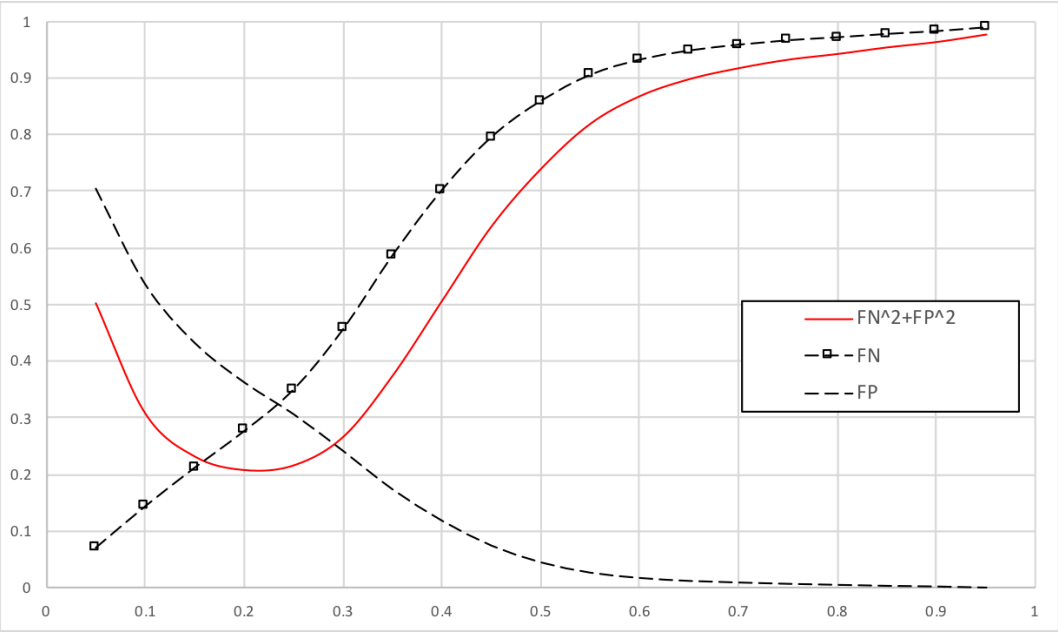
〈그림 3.2〉 문턱값에 따른 모형별 오류비율과 그 제곱합

(a) 모형 M1B, M2B



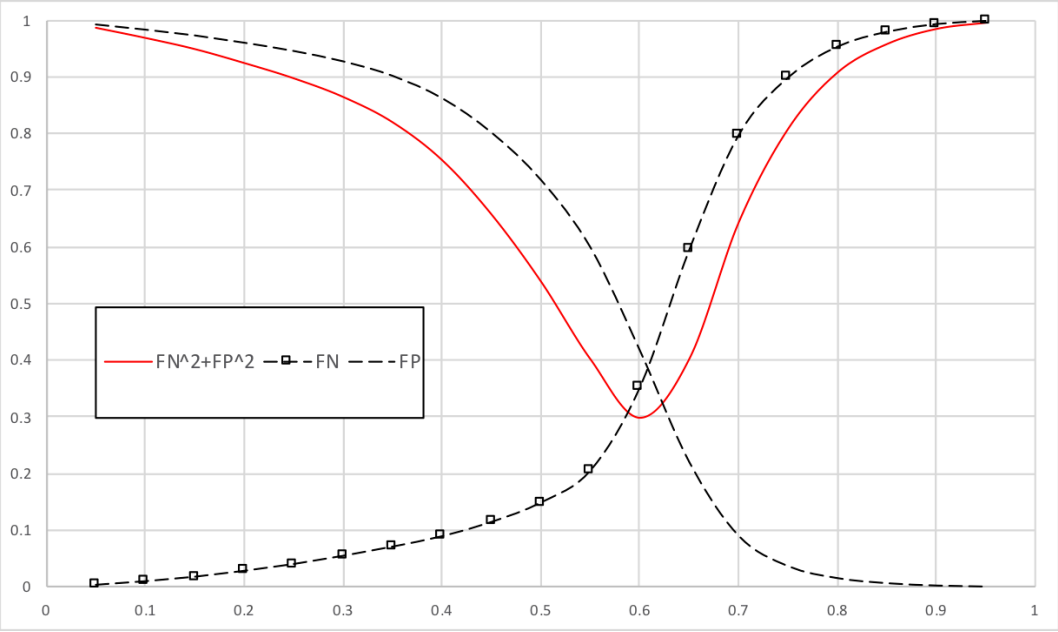
주. FN은 체납건 중 완납으로 잘못 예측한 비율, FP는 완납건 중 체납으로 잘못 예측한 비율, $\text{FN}^2 + \text{FP}^2$은 이 두 비율의 제곱합

(b) 모형 M3B, M4B



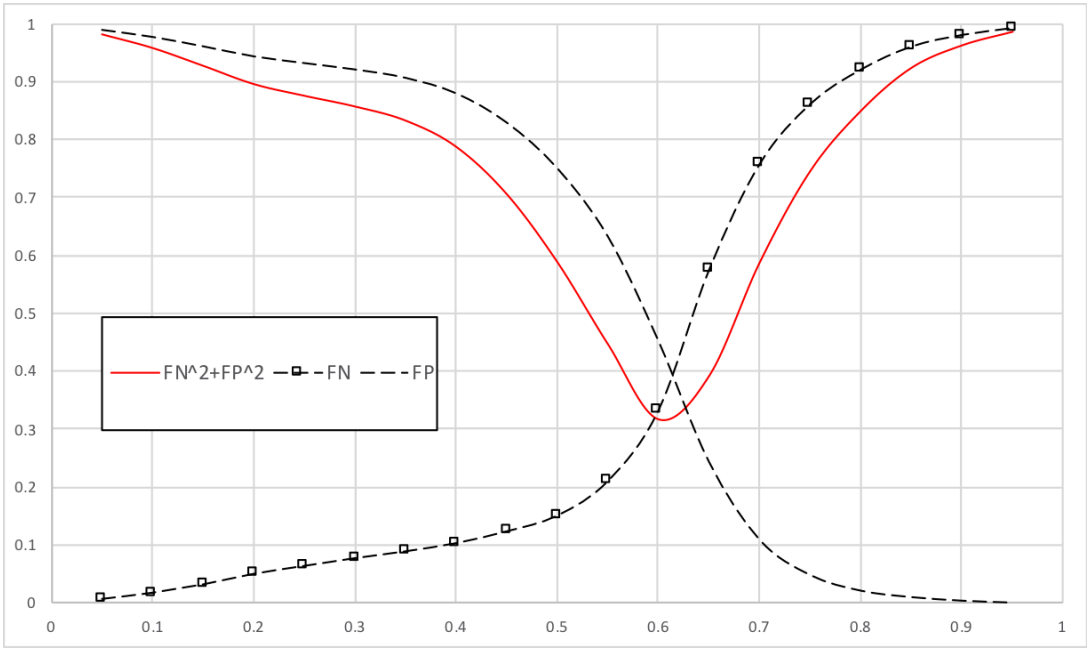
주. FN은 체납건 중 완납으로 잘못 예측한 비율, FP는 완납건 중 체납으로 잘못 예측한 비율, FN^2+FP^2은 이 두 비율의 제곱합

(c) 모형 M5B



주. FN은 체납건 중 완납으로 잘못 예측한 비율, FP는 완납건 중 체납으로 잘못 예측한 비율, FN^2+FP^2은 이 두 비율의 제곱합

(d) 모형 M6B



주. FN은 체납건 중 완납으로 잘못 예측한 비율, FP는 완납건 중 체납으로 잘못 예측한 비율, $FN^2 + FP^2$은 이 두 비율의 제곱합

› 목적함수 값을 최소화하는 p값은 모형에 따라서도 다를 수 있지만 같은 모형이라도 검증용 데이터에 따라 달라질 수 있음. 따라서 정확한 극소화값을 찾기보다 추정에 사용할 수 있는 대략적인 값을 찾고자 하였음

› 여러 차례 무작위로 추출한 검증용 데이터로부터 얻은 $\mathrm{FP}(p)^2 + \mathrm{FN}(p)^2$의 평균값들을 작게 만드는 p값을 모형별 문턱값으로 선정하였음.

⟨표 3.8⟩ 선정된 문턱값과 해당 문턱값에서의 $FP^2 + FN^2$

모형	M1A	M1B	M2A	M2B
문턱값	0.3	0.3	0.3	0.3
$FP^2 + FN^2$	0.2278	0.2106	0.2278	0.2106
모형	M3A	M3B	M4A	M4B
문턱값	0.2	0.2	0.2	0.2
$FP^2 + FN^2$	0.2288	0.2097	0.2288	0.2097
모형	M5A	M5B	M6A	M6B
문턱값	0.6	0.6	0.6	0.6
$FP^2 + FN^2$	0.3096	0.2984	0.3109	0.3163

주. $FP^2 + FN^2$의 값이 작을수록 체납여부 예측성과가 좋은 모형임을 의미함.

› FP^2과 FN^2의 가중합을 이용할 수도 있음

› 체납 데이터와 기한 내 완납 데이터의 비중에 차이가 있고, 체납자들에 대한 체납 예측이 중요한 목적이므로 단순합을 이용하였음

› 전반적으로 2단 모형(M1A~M4B)이 단일모형(M5A~M6B)에 비해 체납 여부에 대한 예측은 다소 양호한 것으로 보임

○ 다음으로 체납된 건들만을 대상으로 체납기간 예측값의 평균제곱오차제곱근(root mean squared errors, RMSE)을 구해 모형의 성과를 비교하였음. 이 값이 작을수록 체납기간의 예측이 정확함을 의미.

$$\text{RMSE}_D = \sqrt{\frac{\sum_j y_j (D_j^C - \hat{D}_j^C)^2}{\sum_j y_j}}. \tag{3.45}$$

› 여러 차례 추출한 검증용 데이터로부터 얻은 RMSE_D값들을 모형별로 평균을 구하면 다음과 같음

⟨표 3.9⟩ 모형별 평균 체납기간 RMSE

모형	M1A	M1B	M2A	M2B
$RMSE_D$	233.7537	253.5113	499.4739	509.7130
모형	M3A	M3B	M4A	M4B
$RMSE_D$	233.7537	253.5113	499.4739	509.7130
모형	M5A	M5B	M6A	M6B
$RMSE_D$	3842.4445	38381419.7618	156.3818	157.0806

주. 체납된 납세고지에 대하여 실제 체납기 간과 예측값 간 차이 제곱의 평균의 제곱근(RMSE)을 의미하며, 이 값이 작을수록 예측성과가 좋음을 의미함

› 모형 M6A와 M6B가 가장 작은 RMSE 값을 보여 줌

› 모형 M3A와 M3B가 그 다음으로 작은 RMSE 값을 보여 줌. 이 두 모형과 M6A와 M6B에 의한 예측의 RMSE 차이는 약 80∼100으로, M6A와 M6B의 예측에 비해 M3A와 M3B의 예측의 오차가 훨씬 큼

○ 종합적 판단: 모형별로 체납 여부 예측의 성과와 체납시 체납기간 예측의 성과를 하나의 표에 나타내 보면 다음과 같음.

⟨표 3.10⟩ 모형별 평균 종합적 예측 성과

모형	M1A	M1B	M2A	M2B
체납여부 예측오차	0.2278 (0.3)	0.2106 (0.3)	0.2278 (0.3)	0.2106 (0.3)
체납기간 예측오차	233.7537	253.5113	499.4739	503.7130
모형	M3A	M3B	M4A	M4B
체납여부 예측오차	0.2288 (0.2)	0.2097 (0.2)	0.2288 (0.2)	0.2097 (0.2)
체납기간 예측오차	233.7537	253.5113	499.4739	503.7130
모형	M5A	M5B	M6A	M6B
체납여부 예측오차	0.3096 (0.6)	0.2984 (0.6)	0.3109 (0.6)	0.3163 (0.6)
체납기간 예측오차	3842.4445	38381419.7618	156.3818	157.0806

출처: 국세청 제공 데이터로부터 자체 계산

주: '체납여부 예측오차'는 $FN^2 + FP^2$이며, '체납기간 예측오차'는 체납된 납세고지에 대하여 실제 체납기간과 예측값 간 차이 제곱의 평균의 제곱근(RMSE)을 의미함. 두 지표 모두 작을수록 좋은 모형임을 의미함. '체납여부 예측오차'의 괄호 안의 값은 사용한 최적 문턱값.

› 체납 여부 예측 성과 면에서는 M3B와 M4B(두 모형은 동일한 체납여부 확률 추정 모형임)가 가장 우수함

› 체납시 체납 기간 예측 측면에서 M6A와 M6B가 다른 모형에 비해 확연히 우수한 성과를 보여 주고 있음

› 체납여부를 판단하고자 할 경우 M3B와 M4B가 적절함

› 종합적인 측면에서 M6A와 M6B를 판단할 때, M6A가 M6B보다 근소하게 더 나으나 M6A에서 이용하는 비율형 재무제표 변수들에서 분모가 0이 되는 경우도 많았고 변수들의 값의 분산이 매우 커서 로그형(M6B)을 이용하는 것이 더 적절한 것으로 판단됨

○ 추정에 이용된 M6B모형에 대해 x와 납세 기한$(T - \tau)$이 체납 확률과 체납 정리기간에 미치는 한계적 영향을 도출해 보면 다음과 같음

$$\frac{d\Pr[y=1|x,\tau,T]}{dx} = \phi(\ln\alpha)\,\frac{\beta}{\sigma}, \tag{3.46}$$

$$\frac{d\Pr[y=1|x,\tau,T]}{d(T-\tau)} = -\frac{1}{(T-\tau)\sigma}\phi(\ln\alpha), \tag{3.47}$$

$$\frac{d\ln(E[D^C|x,\tau,T,y=1])}{dx} = \frac{\beta}{\sigma}\left[\sigma + \frac{\phi(\ln\alpha+\sigma)}{\Phi(\ln\alpha+\sigma)} - \frac{\phi(\ln\alpha)}{\Phi(\ln\alpha)}\right], \tag{3.48}$$

$$\frac{d\ln(E[D^C|x,\tau,T,y=1])}{d(T-\tau)} = -\frac{1}{\sigma(T-\tau)}\left[\frac{\phi(\ln\alpha+\sigma)}{\Phi(\ln\alpha+\sigma)} - \frac{\phi(\ln\alpha)}{\Phi(\ln\alpha)}\right], \tag{3.49}$$

$$\ln\alpha = \frac{1}{\sigma}[\ln(T-\tau) - x'\beta]$$

○ 식 (3.46)은 설명변수 x가 체납확률에 미치는 한계 효과의 방향은 x의 계수의 부호와 같음을 의미함

○ 식 (3.47)은, 다른 조건이 동일할 때, 납세 기한$(T-\tau)$이 길수록 체납확률이 낮아짐을 의미함

○ 식 (3.48)의 우변에서 $\sigma + \dfrac{\phi(A+\sigma)}{\Phi(A+\sigma)} - \dfrac{\phi(A)}{\Phi(A)} > 0$이므로, 식 (3.48)은 설명변수 x가 평균 체납정리 기간에 미치는 한계 효과의 방향이 x의 계수의 부호와 같음을 의미함

○ 식 (3.49)에서, $\dfrac{\phi(A+\sigma)}{\Phi(A+\sigma)} - \dfrac{\phi(A)}{\Phi(A)} < 0$이므로, 식 (3.49)는, 다른 조건이 동일할 때, 납세 기한$(T-\tau)$이 길어질 경우 평균 체납정리기간도 길어짐을 의미함

3.4.6 모형 추정 결과

□ 위의 결정방법에 따라 M6B를 최종 선정함

□ 전체 데이터를 이용해 M6B의 모수들을 추정한 결과는 ⟨표 3.3⟩에 제시하였음

○ 필요시 참고할 수 있도록 M1B ∼ M5B에 대한 추정 결과도 부록에 제시함

□ 추정 결과에 대한 설명 및 해석은 3.3절 참고

3.4.7 부록 A: 모형별 추정 결과

〈표 3.11〉 M1B 추정 결과

체납 확률 모형		Model = M1B	체납 기간 모형	
계수	t-값	변수명	계수	t-값
0.8434761	7.807	상수항	3.652426	22.68885
-0.1494445	-9.298	추세	0.401898	16.46279
-0.0238609	-34.124	연령	-0.024886	-24.14709
-0.0182225	-3.535	세대원수	-0.042958	-6.18026
-0.2918786	-19.838	자동차보유여부	-0.325004	-15.91396
-0.1990027	-1.772	명의위장사업자여부	0.304977	1.98795
-0.2035221	-3.224	전년도유예신청건수	-0.374995	-4.82478
-0.4969481	-3.74	직전2년평균유예신청건수	-0.062196	-0.38298
1.4013416	10.344	직전3년평균유예신청건수	-0.012433	-0.07814
0.2025041	63.088	전년도고지서개수	0.042401	11.3547
0.0405028	5.939	ln차감고지세액	0.203092	20.67297
-0.7171646	-40.865	ln전년도종합소득	0.034403	1.29926
1.5648055	14.197	ln전년도유동자산금액	0.842768	5.21717
-0.6773863	-6.849	ln전년도당좌자산금액	-0.556771	-3.71604
0.1099482	9.594	ln전년도매출채권금액	0.202234	11.9097
-0.1656079	-16.486	ln전년도재고자산금액	-0.143151	-9.57915
-1.2506838	-16.41	ln전년도자산총계금액	-0.527632	-4.61292
0.3961494	8.686	ln전년도유동부채금액	0.307278	6.00857
0.0245075	2.243	ln전년도매입채무금액	-0.010111	-0.62228

체납 확률 모형		Model = M1B	체납 기간 모형	
계수	t-값	변수명	계수	t-값
0.0547849	5.391	ln전년도단기차입금액	-0.061309	-4.20917
0.0080994	1.164	ln전년도유동성장기부채금액	0.009035	0.97125
0.1083506	8.69	ln전년도장기차입금액	0.007236	0.39786
-0.0009464	-0.123	ln전년도선수금액	-0.00705	-0.61407
-0.0150408	-1.888	ln전년도선급금액	-0.005177	-0.4371
-0.480688	-9.174	ln전년도부채총계금액	-0.355787	-5.49274
-0.3046973	-14.586	ln전년도자본총계금액	-0.29601	-9.57043
1.525023	16.262	ln전년도매출금액	0.96217	8.81543
0.2894587	1.532	ln전년도영업이익금액	0.000354	0.00149
0.0144074	0.262	ln전년도영업손실금액	-0.02646	-0.39064
-0.1055666	-12.143	ln전년도현금성자산금액	-0.099513	-7.24942
-1.0870149	-6.033	ln전년도당기순이익금액	-0.412763	-1.84145
-0.2998498	-5.34	ln전년도당기순손실금액	-0.065192	-0.96135
-0.0926704	-6.352	ln전년도이자비용금액	-0.130084	-6.30278
-0.1039888	-12.012	ln전년도이자수익금액	-0.142615	-10.49872
		$\ln(\sigma)$	0.541993	128.32519
		σ	1.72	
		Loglike(model)	-189553.9	
		Loglike(const.)	-192086.6	
137,234		관측치 수	36,896	

〈표 3.12〉 M2B 추정 결과

체납확률모형		Model = M2B	체납기간모형	
계수	t-값	변수명	계수	t-값
0.8434761	7.807	상수항	2.35621	14.274
-0.1494445	-9.298	추세	0.50586	20.369
-0.0238609	-34.124	연령	-0.02874	-27.407
-0.0182225	-3.535	세대원수	-0.04538	-6.116
-0.2918786	-19.838	자동차보유여부	-0.30855	-14.54
-0.1990027	-1.772	명의위장사업자여부	0.17968	1.175
-0.2035221	-3.224	전년도유예신청건수	-0.43251	-5.248
-0.4969481	-3.74	직전2년평균유예신청건수	-0.04633	-0.273
1.4013416	10.344	직전3년평균유예신청건수	0.12589	0.747
0.2025041	63.088	전년도고지서개수	0.05312	15.31
0.0405028	5.939	ln차감고지세액	0.11445	11.32
-0.7171646	-40.865	ln전년도종합소득	0.05987	2.184
1.5648055	14.197	ln전년도유동자산금액	1.09464	6.573
-0.6773863	-6.849	ln전년도당좌자산금액	-0.69992	-4.553
0.1099482	9.594	ln전년도매출채권금액	0.1743	10.144
-0.1656079	-16.486	ln전년도재고자산금액	-0.1578	-10.334
-1.2506838	-16.41	ln전년도자산총계금액	-0.5477	-4.966
0.3961494	8.686	ln전년도유동부채금액	0.32208	5.468
0.0245075	2.243	ln전년도매입채무금액	0.0076	0.461

체납확률모형		Model = M2B	체납기간모형	
형 계수	t-값	변수명	계수	t-값
0.0547849	5.391	ln전년도단기차입금액	-0.05839	-3.868
0.0080994	1.164	ln전년도유동성장기부채금액	0.00633	0.669
0.1083506	8.69	ln전년도장기차입금액	0.00964	0.511
-0.0009464	-0.123	ln전년도선수금액	0.00178	0.153
-0.0150408	-1.888	ln전년도선급금액	-0.02142	-1.761
-0.480688	-9.174	ln전년도부채총계금액	-0.3163	-4.488
-0.3046973	-14.586	ln전년도자본총계금액	-0.28496	-9.523
1.525023	16.262	ln전년도매출금액	1.09344	9.428
0.2894587	1.532	ln전년도영업이익금액	-0.08592	-0.351
0.0144074	0.262	ln전년도영업손실금액	-0.04609	-0.657
-0.1055666	-12.143	ln전년도현금성자산금액	-0.09733	-6.9
-1.0870149	-6.033	ln전년도당기순이익금액	-0.46109	-2.031
-0.2998498	-5.34	ln전년도당기순손실금액	-0.09091	-1.297
-0.0926704	-6.352	ln전년도이자비용금액	-0.14463	-6.759
-0.1039888	-12.012	ln전년도이자수익금액	-0.14351	-10.376
		$\ln(\sigma)$	0.61715	158.536
		σ	1.85	
		Loglike(model)	-186923.7	
		Loglike(const.)	-190759.7	
137,234		관측치 수	36,896	

〈표 3.13〉 M3B 추정 결과

체납확률모형		Model = M3B	체납기간모형	
형 계수	t-값	변수명	계수	t-값
0.8434761	7.807	상수항	3.652426	22.68885
-0.1494445	-9.298	추세	0.401898	16.46279
-0.0238609	-34.124	연령	-0.024886	-24.14709
-0.0182225	-3.535	세대원수	-0.042958	-6.18026
-0.2918786	-19.838	자동차보유여부	-0.325004	-15.91396
-0.1990027	-1.772	명의위장사업자여부	0.304977	1.98795
-0.2035221	-3.224	전년도유예신청건수	-0.374995	-4.82478
-0.4969481	-3.74	직전2년평균유예신청건수	-0.062196	-0.38298
1.4013416	10.344	직전3년평균유예신청건수	-0.012433	-0.07814
0.2025041	63.088	전년도고지서개수	0.042401	11.3547
0.0405028	5.939	ln차감고지세액	0.203092	20.67297
-0.7171646	-40.865	ln전년도종합소득	0.034403	1.29926
1.5648055	14.197	ln전년도유동자산금액	0.842768	5.21717
-0.6773863	-6.849	ln전년도당좌자산금액	-0.556771	-3.71604
0.1099482	9.594	ln전년도매출채권금액	0.202234	11.9097
-0.1656079	-16.486	ln전년도재고자산금액	-0.143151	-9.57915
-1.2506838	-16.41	ln전년도자산총계금액	-0.527632	-4.61292
0.3961494	8.686	ln전년도유동부채금액	0.307278	6.00857
0.0245075	2.243	ln전년도매입채무금액	-0.010111	-0.62228

체납확률모형		Model = M3B	체납기간모형	
형 계수	t-값	변수명	계수	t-값
0.0547849	5.391	ln전년도단기차입금액	-0.061309	-4.20917
0.0080994	1.164	ln전년도유동성장기부채금액	0.009035	0.97125
0.1083506	8.69	ln전년도장기차입금액	0.007236	0.39786
-0.0009464	-0.123	ln전년도선수금액	-0.00705	-0.61407
-0.0150408	-1.888	ln전년도선급금액	-0.005177	-0.4371
-0.480688	-9.174	ln전년도부채총계금액	-0.355787	-5.49274
-0.3046973	-14.586	ln전년도자본총계금액	-0.29601	-9.57043
1.525023	16.262	ln전년도매출금액	0.96217	8.81543
0.2894587	1.532	ln전년도영업이익금액	0.000354	0.00149
0.0144074	0.262	ln전년도영업손실금액	-0.02646	-0.39064
-0.1055666	-12.143	ln전년도현금성자산금액	-0.099513	-7.24942
-1.0870149	-6.033	ln전년도당기순이익금액	-0.412763	-1.84145
-0.2998498	-5.34	ln전년도당기순손실금액	-0.065192	-0.96135
-0.0926704	-6.352	ln전년도이자비용금액	-0.130084	-6.30278
-0.1039888	-12.012	ln전년도이자수익금액	-0.142615	-10.49872
		$\ln(\sigma)$	0.541993	128.32519
		σ	1.72	
		Loglike(model)	-189553.9	
		Loglike(const.)	-192086.6	
137,234		관측치 수	36,896	

〈표 3.14〉 M4B 추정 결과

체납확률모형		Model = M4B	체납기간모형	
형 계수	t-값	변수명	계수	t-값
0.8434761	7.807	상수항	2.35621	14.274
-0.1494445	-9.298	추세	0.50586	20.369
-0.0238609	-34.124	연령	-0.02874	-27.407
-0.0182225	-3.535	세대원수	-0.04538	-6.116
-0.2918786	-19.838	자동차보유여부	-0.30855	-14.54
-0.1990027	-1.772	명의위장사업자여부	0.17968	1.175
-0.2035221	-3.224	전년도유예신청건수	-0.43251	-5.248
-0.4969481	-3.74	직전2년평균유예신청건수	-0.04633	-0.273
1.4013416	10.344	직전3년평균유예신청건수	0.12589	0.747
0.2025041	63.088	전년도고지서개수	0.05312	15.31
0.0405028	5.939	ln차감고지세액	0.11445	11.32
-0.7171646	-40.865	ln전년도종합소득	0.05987	2.184
1.5648055	14.197	ln전년도유동자산금액	1.09464	6.573
-0.6773863	-6.849	ln전년도당좌자산금액	-0.69992	-4.553
0.1099482	9.594	ln전년도매출채권금액	0.1743	10.144
-0.1656079	-16.486	ln전년도재고자산금액	-0.1578	-10.334
-1.2506838	-16.41	ln전년도자산총계금액	-0.5477	-4.966
0.3961494	8.686	ln전년도유동부채금액	0.32208	5.468
0.0245075	2.243	ln전년도매입채무금액	0.0076	0.461

체납확률모형		Model = M4B	체납기간모형	
형 계수	t-값	변수명	계수	t-값
0.0547849	5.391	ln전년도단기차입금액	-0.05839	-3.868
0.0080994	1.164	ln전년도유동성장기부채금액	0.00633	0.669
0.1083506	8.69	ln전년도장기차입금액	0.00964	0.511
-0.0009464	-0.123	ln전년도선수금액	0.00178	0.153
-0.0150408	-1.888	ln전년도선급금액	-0.02142	-1.761
-0.480688	-9.174	ln전년도부채총계금액	-0.3163	-4.488
-0.3046973	-14.586	ln전년도자본총계금액	-0.28496	-9.523
1.525023	16.262	ln전년도매출금액	1.09344	9.428
0.2894587	1.532	ln전년도영업이익금액	-0.08592	-0.351
0.0144074	0.262	ln전년도영업손실금액	-0.04609	-0.657
-0.1055666	-12.143	ln전년도현금성자산금액	-0.09733	-6.9
-1.0870149	-6.033	ln전년도당기순이익금액	-0.46109	-2.031
-0.2998498	-5.34	ln전년도당기순손실금액	-0.09091	-1.297
-0.0926704	-6.352	ln전년도이자비용금액	-0.14463	-6.759
-0.1039888	-12.012	ln전년도이자수익금액	-0.14351	-10.376
		$\ln(\sigma)$	0.61715	158.536
		σ	1.85	
		Loglike(model)	-186923.7	
		Loglike(const.)	-190759.7	
137,234		관측치 수	36,896	

〈표 3.15〉 M5B, M6B 추정 결과

Model = M5B			Model = M6B	
계수	t-값	변수명	계수	t-값
3.57075	74.97	상수항	3.11772	78.0979
3.57075	74.97	상수항	3.11772	78.0979
0.08527	11.112	추세	0.08988	14.3979
-0.00212	-33.108	연령	0.09263	32.9708
-0.00457	-1.845	세대원수	-0.00343	-21.9541
-0.25874	-33.972	자동차보유여부	-0.00388	-1.8482
0.34997	5.942	명의위장사업자여부	-0.16082	-25.7103
-0.1089	-3.415	전년도유예신청건수	0.09304	1.9567
-0.11948	-1.753	직전2년평균유예신청건수	-0.23138	-33.2733
0.68853	9.841	직전3년평균유예신청건수	-0.09655	-3.7293
0.10242	62.074	전년도고지서개수	-0.10256	-1.8267
0.15687	46.092	ln차감고지세액	0.70183	12.1145
-0.31166	-34.882	ln전년도종합소득	0.06727	59.5218
1.06581	19.223	ln전년도유동자산금액	0.71467	16.225
-0.67799	-12.71	ln전년도당좌자산금액	-0.44184	-10.5822
0.12446	23.2	ln전년도매출채권금액	0.07343	16.7537
-0.10715	-22.302	ln전년도재고자산금액	-0.07148	-18.0844
-0.57793	-17.313	ln전년도자산총계금액	-0.32634	-12.2987
0.1545	10.952	ln전년도유동부채금액	0.09369	7.2743
-0.00294	-0.552	ln전년도매입채무금액	0.0021	0.4833

Model = M5B			Model = M6B	
계수	t-값	변수명	계수	t-값
0.0057	1.208	ln전년도단기차입금액	0.01901	4.8427
0.00751	2.211	ln전년도유동성장기부채금액	0.00593	2.1666
0.05061	8.844	ln전년도장기차입금액	0.0433	9.1073
-0.00871	-2.464	ln전년도선수금액	-0.00443	-1.5317
-0.00499	-1.383	ln전년도선급금액	-0.01042	-3.5011
-0.26109	-14.413	ln전년도부채총계금액	-0.14044	-8.8263
-0.24949	-22.041	ln전년도자본총계금액	-0.1582	-18.2015
0.63044	25.283	ln전년도매출금액	0.4152	17.9463
-0.02642	-0.405	ln전년도영업이익금액	0.00306	0.0508
-0.03035	-1.644	ln전년도영업손실금액	-0.00393	-0.2269
-0.06715	-18.015	ln전년도현금성자산금액	-0.03562	-11.7405
-0.07071	-1.13	ln전년도당기순이익금액	-0.14314	-2.5036
0.01972	1.032	ln전년도당기순손실금액	-0.03471	-1.9596
-0.08168	-12.127	ln전년도이자비용금액	-0.04645	-8.2313
-0.08067	-22.038	ln전년도이자수익금액	-0.04427	-15.0212
0.1501	77.344	$\ln(\sigma)$	-0.03262	-16.0673
1.16		σ	0.968	
-632093.6		Loglike(model)	-603748.7	
-640620.7		Loglike(const.)	-609312.4	
126,086		관측치 수	126,086	

3.4.8 부록 B: 표준화에 이용한 a와 b 및 변수들의 평균

□ $(x-a)/b$ 표준화에 사용된 a와 b

⟨표 3.16⟩ 분석에 사용된 데이터 전체의 평균과 a, b

변수명	평균	a	b
연령	54.02	0	1
세대원수	3.262	0	1
자동차보유여부	0.7408	0	1
명의위장사업자여부	0.003181	0	1
전년도유예신청건수	0.02538	0	1
직전2년평균유예신청건수	0.02257	0	1
직전3년평균유예신청건수	0.02094	0	1
전년도고지서개수	2.832	0	1
ln차감고지세액	16.17	16.1793	0.64922
ln전년도종합소득	12.42	12.487	8.651988
전년도부채비율	2.164×10^{10}	2.258296×10^{10}	3.171989×10^{11}
전년도유동부채비율	7.344×10^{10}	7.796588×10^{9}	1.142766×10^{11}
전년도차입금의존도	2.353×10^{6}	2.299926×10^{6}	3.704029×10^{8}
전년도단기성차입금비율	-1.365×10^{5}	-1.334350×10^{5}	2.947179×10^{7}
전년도재고자산유동자산비율	0.069946	0.07000412	0.1903997
전년도당좌비율	1.765×10^{9}	1.771111×10^{9}	3.337619×10^{10}
전년도현금성자산비율	3.306×10^{9}	3.268567×10^{9}	5.720923×10^{10}
전년도영업이익률	-4.331×10^{7}	-4.257948×10^{7}	1.959309×10^{9}
전년도매출액순이익률	-6.841×10^{17}	-6.710961×10^{7}	3.676641×10^{9}
전년도금융비용매출액비율	2.936×10^{7}	2.868946×10^{7}	1.816597×10^{9}
전년도재고자산회전율	2.076×10^{11}	2.102775×10^{11}	5.490144×10^{11}
전년도매출채권회전율	2.078×10^{11}	2.069331×10^{11}	8.335470×10^{11}
전년도매입채무회전율	2.204×10^{11}	2.207339×10^{11}	6.269102×10^{11}
전년도운전자금회전율	7.657×10^{10}	7.733165×10^{10}	2.651635×10^{11}

	평균	a	b
ln전년도유동자산금액	11.5	11.56375	8.760095
ln전년도당좌자산금액	11.35	11.41696	8.665460
ln전년도매출채권금액	5.914	5.991753	8.637665
ln전년도재고자산금액	4.567	4.565497	7.475824
ln전년도자산총계금액	12.64	12.70128	9.415438
ln전년도유동부채금액	10.63	10.69820	8.553179
ln전년도매입채무금액	4.863	4.905709	7.814049
ln전년도단기차입금액	2.875	2.869655	6.761205
ln전년도유동성장기부채금액	0.01237	0.01262155	0.4766318
ln전년도장기차입금액	4.191	4.186955	8.035116
ln전년도선수금액	0.4096	0.4120615	2.572485
ln전년도선급금액	0.6438	0.6451851	3.158802
ln전년도부채총계금액	11.55	11.60143	9.029249
ln전년도자본총계금액	11.65	11.68359	9.228131
ln전년도매출금액	12.87	12.94772	9.637114
ln전년도영업이익금액	11.33	11.39697	8.824375
ln전년도영업손실금액	0.369	0.3727761	2.463420
ln전년도현금성자산금액	1.104	1.087799	4.194288
ln전년도당기순이익금액	11.2	11.26411	8.786536
ln전년도당기순손실금액	0.4298	0.4341788	2.655340
ln전년도이자비용금액	5.202	5.194144	7.544984
ln전년도이자수익금액	0.5972	0.5905074	2.467988

□ 체납의 경우와 체납기록이 없는 경우의 평균값

〈표 3.17〉 체납인 경우와 미체납인 경우의 평균

체납이 아닌 경우	변수명	체납인 경우
54.71	연령	52.22
3.252	세대원수	3.288
0.7575	자동차보유여부	0.697
0.002971	명의위장사업자여부	0.003732
0.02272	전년도유예신청건수	0.03238
0.02005	직전2년평균유예신청건수	0.02916
0.01821	직전3년평균유예신청건수	0.0281
2.813	전년도고지서개수	2.881
16.18	ln차감고지세액	16.16
13.57	ln전년도종합소득	9.414
2.195×10^{10}	전년도부채비율	2.085×10^{10}
6.83×10^{9}	전년도유동부채비율	8.691×10^{9}
1.682×10^{6}	전년도차입금의존도	4113000
-27230	전년도단기성차입금비율	-423000
0.07488	전년도재고자산유동자산비율	0.05702
1.988×10^{9}	전년도당좌비율	1.18×10^{9}
4.01×10^{9}	전년도현금성자산비율	1.459×10^{9}
-5.491×10^{7}	전년도영업이익률	-1.289×10^{7}
-8.622×10^{7}	전년도매출액순이익률	-2.174×10^{7}
3.815×10^{7}	전년도금융비용매출액비율	6.301×10^{6}
2.185×10^{11}	전년도재고자산회전율	1.792×10^{11}
2.325×10^{11}	전년도매출채권회전율	1.43×10^{11}
2.425×10^{11}	전년도매입채무회전율	1.626×10^{11}
8.297×10^{10}	전년도운전자금회전율	5.98×10^{10}

체납이 아닌 경우	변수명	체납인 경우
12.4	ln전년도유동자산금액	9.137
12.24	ln전년도당좌자산금액	9.014
6.175	ln전년도매출채권금액	5.229
4.981	ln전년도재고자산금액	3.482
13.73	ln전년도자산총계금액	9.793
11.43	ln전년도유동부채금액	8.538
5.143	ln전년도매입채무금액	4.132
3.081	ln전년도단기차입금액	2.334
0.01199	ln전년도유동성장기부채금액	0.01338
4.651	ln전년도장기차입금액	3.221
0.4492	ln전년도선수금액	0.3059
0.7087	ln전년도선급금액	0.4736
12.5	ln전년도부채총계금액	9.037
12.74	ln전년도자본총계금액	8.809
13.87	ln전년도매출금액	10.23
12.29	ln전년도영업이익금액	8.835
0.382	ln전년도영업손실금액	0.335
1.279	ln전년도현금성자산금액	0.6437
12.15	ln전년도당기순이익금액	8.712
0.4409	ln전년도당기순손실금액	0.4008
5.678	ln전년도이자비용금액	3.954
0.6824	ln전년도이자수익금액	0.3741

제 4 장

결론

□ 본 연구의 목적은 체납 확률 및 체납이 발생하는 경우 체납 기간을 추정하는 모형을 제시하고, 국세청 제공 기초자료를 가공하여 분석에 합당한 데이터를 생성하며, 이렇게 생성된 데이터를 이용하여 분석하는 것임

○ 모형을 통해 예측되는 체납확률과 체납기간을 납세자 성향 파악에 이용할 수 있음

□ 2장에서는 국세청 기초 자료를 분석 목적에 맞도록 가공하는 방법을 제시하고 국세청 제공 자료를 이에 따라 가공하며, 3장에서는 이 자료를 이용하여 체납 정리기간 결정요인을 추정하는 모형을 선정하고 추정함

○ 2장에서는 기초자료의 형태로 저장된 다양한 데이터베이스를 가공하여 하나의 분석 가능한 데이터셋을 갖추는 방법을 R 소스 코드와 함께 제시하였음

› 이와 더불어, 국세청 업무담당자와 연구자의 효율적인 업무분담이 가능한 분업 모델을 제시하고 활용하였음

○ 3장에서는 12개의 체납확률 예측모형을 비교하여 최적의 모형을 선별하고, 이를 이용한 추정결과를 제시하였음

› 분석 결과에 의하면, 고려된 12개 모형 중 로그정규분포를 사용한 단일(one-tier) 모형에 재무제표상 변수값으로 비율보다는 로그값을 사용한 모형이 최적인 것으로 판명되었음

› 단, 이 결과는 본 연구를 위하여 국세청에서 제공한 특정 자료에 한하며, 추후 확장된 데이터를 분석할 경우 보다 강력한 기계학습 등의 방법을 사용하여 체납여부 및 체납 시 정리기간을 예측하는 연구가 필요함

○ 연구 결과에 의하면, 여타 요소가 동일할 때 체납 확률은 증가 추세에 있음

○ 개인성향변수, 납세이력변수, 재무제표변수 등을 이용해 분석했고 세 가지 유형의 변수들 모두 전반적으로 유의한 것으로 나타났음(상세한 결과는 〈표 3.3〉 참조)

□ 정책적 제언: 추정 결과들을 바탕으로 개별 납세자의 체납 가능성과 체납 시 정리기간을 예측하는 데에 활용할 수 있음

○ 본 연구는 국세청 데이터의 일부만을 사용하여 진행되었으며, 향후 체납 가능성과 체납 시 정리기간 예측을 위한 데이터 기반 모형의 개발 시에 지침으로 활용될 수 있음

○ 방대한 국세청 데이터와 추가적 변수 활용 시 예측성과는 보다 향상될 것으로 기대함

□ 국세청 데이터의 성격상 엄격한 보안을 유지하면서 효율적인 분석을 수행하는 것이 쉽지 않다는 난점이 있음에도 국세청 본청과 서울지방국세청 간의 긴밀한 협력체계를 이용하여 결과를 도출할 수 있었음

○ 국세청 데이터는 크기가 매우 방대하며, 분석의 성격상 다양한 모형의 검토가 필요하여, 보안을 유지하면서 효율적인 분석을 수행하는 것이 쉽지 않다는 난점이 있음

○ 본 연구에서는 이러한 제약하에서, 국세청 본청과 서울지방국세청 간의 긴밀한 협력체계를 이용하여 방대한 작업을 수행하고 결과를 도출할 수 있었음

○ 이는 일면 데이터 범위를 제한시켜 분석이 가능한 정도의 크기로 한정하였기 때문이며, 더욱 방대한 데이터를 기계학습 등 막강한 도구로써 분석하기 위해서는 보안을 유지하면서도 분석의 효율성을 높일 수 있는 효과적인 방안의 마련이 전제되어야 할 것으로 보임

참고문헌

Borgia, Carl R., Philip H. Siegel, and Dennis Ortiz (2014). A survival analysis of tax professionals' performance and internship experience. Accounting Research Journal, Emerald Group Publishing, vol. 27(3), 266–285.

Garson, G. D. (2012). Parametric Survival Analysis. Asheboro, NC: Statistical Associates Publishers.

Lambert, Paul C., and Patrick Royston, 2009. Further development of flexible parametric models for survival analysis. Stata Journal, StataCorp LP, vol. 9(2), 265–290.

Miller, Rupert G. (1997), Survival analysis, John Wiley & Sons.

OECD (2014). Working Smarter in Tax Debt Management. OECD Publishing.

Richards, S. J. (2012). A handbook of parametric survival models for actuarial use. Scandinavian Actuarial Journal. 2012 (4): 233–257.

Wooldridge, Jeffrey M. (2002). Econometric Analysis of Cross Section and Panel Data, MIT Press.

한국조세연구원(2010). 체납정리 인프라 개선방안. 연구용역보고서.

한국조세재정연구원(2018). 빅데이터와 조세행정 : 최근 해외 트렌드를 중심으로. 연구용역보고서.