강의 계획 및 평가 방법

주차	날짜	강의 내용	과제 주제	대면/비대면	평가
1	03/06	강의 소개		Online	
2	03/13	데이터 마이닝 절차		A704	
3	03/20	데이터 탐색 및 시각화		B224	
4	03/27	차원 축소	과제 1	Online	
5	04/03	예측성능 평가		Online	과제 1 (10%)
6	04/10	다중 선형 회귀분석		A704	
7	04/17	중간 프로젝트 발표		A704	30%
8	04/24	k-최근접이웃 알고리즘 나이브 베이즈 분류		Online	
9	05/01 보강: 06/15(목)	휴업일(근로자의 날) 동영상 강의		Online	
10	05/08	분류와 회귀 나무	과제 2	Online	
11	**05/15**	**로지스틱 회귀분석**		**A704**	과제 2 (10%)
12	05/22	신경망 판별 분석		Online	
13	05/29 보강: 06/02(금) 19시	대체 공휴일(부처님 오신 날) 연관 규칙		Online	
14	06/05	군집 분석		A704	
15	06/12	기말 프로젝트 발표		B224	40%

Data Mining for Business Analytics

Ch. 10 Logistic Regression

2023.05.15.

Contents

10.1 Introduction

- 로지스틱 회귀: 선형회귀의 개념을 종속변수 Y가 범주형인 경우로 확장
  - ✓ ex) 주식의 보유/매도/매수
  - ✓ 선형회귀의 경우 결과변수 → 연속형 변수
- 로지스틱 회귀의 용도
  - ✓ 분류(Classification): 예측변수의 값을 바탕으로 클래스가 알려져 있지 않은 새로운 관측값의 클래스를 결정

    ex) 고객을 재구매 고객과 처음 구매한 고객으로 분류, 신용점수와 같은 정보로부터 대출의 승인 또는 비승인 예측
  - ✓ 프로파일링(Profiling): 서로 다른 클래스의 관측값들을 구별하는 요인을 찾을 때

    ex) 남자 최고 경영진과 여자 최고 경영진을 구별하는 요인 찾기
- 변수에 대한 가정(본 강의)
  - ✓ 종속변수: 이진형(binary) 변수 (0 or 1) ex) 성공/실패, 구매/비구매 등
  - ✓ 예측변수: 연속형 또는 범주형
- 로지스틱 회귀의 단계

1. 각 클래스에 속하는 확률 추정: $p = P(Y = 1)$ (클래스 '1'에 속할 확률)
2. 각 관측값의 클래스를 지정하기 위해 확률값에 대한 cutoff 값 적용

ex) Let cutoff = 0.5

if $P(Y = 1) > 0.5$, 클래스 '1'로 분류, if $P(Y = 1) < 0.5$, 클래스 '0'로 분류

10.2 The Logistic Regression Model

- 로지스틱 회귀의 원리: 종속변수로 Y를 대신하여 '로짓(logit)'이라고 부르는 Y의 함수 사용. 즉 로짓을 예측변수들의 선형함수로 모델링. 로짓이 예측되면 그로부터 확률 매핑

Logistic Response Function(로지스틱 반응함수)

- 클래스 '1'에 속할 확률 구함 ($p = P(Y = 1)$)
  - ✓ 실제 Y는 '0'과 '1'의 값만 가짐 → 확률로 변환
  - ✓ $\begin{cases} P(Y = 1) = p \\ P(Y = 0) = 1 - p \end{cases}$
  - ✓ 확률 p 를 q 개의 예측변수들에 대한 선형변수로 다음과 같이 나타낸다면

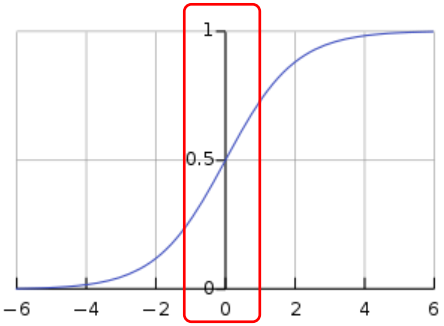
$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q \qquad (-\infty < p < \infty)$$

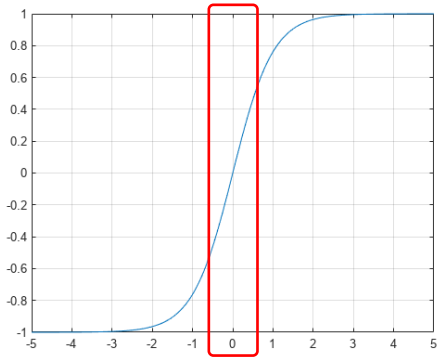
  - ✓ 우변항이 구간 [0, 1]에 들어간다는 것을 보장할 수 없음 → 예측변수들에 대한 비선형 함수 사용
  - ✓ **로지스틱 반응함수(logistic response function)**

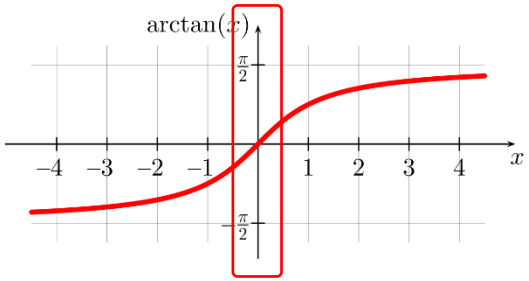
$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q)}} = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q} + 1} \qquad (0 \le p \le 1)$$

10.2 The Logistic Regression Model

- 시그모이드(Sigmoid) 함수
  - ✓ S자형 곡선 또는 시그모이드 곡선을 갖는 수학 함수
  - ✓ 실수 전체를 정의역으로 가지며, 반환값은 단조증가하거나 단조감소함
  - ✓ 함수값은 [0, 1] 또는 [−1, 1]의 범위를 가짐
  - ✓ 인공 뉴런의 활성화 함수로 사용됨
  - ✓ 종류
    - 로지스틱 함수 $$f(x) = \frac{1}{1 + e^{-x}}$$
    - 쌍곡탄젠트(hyperbolic) $$f(x) = \tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$
    - 아크탄젠트 함수 $$f(x) = \arctan x$$

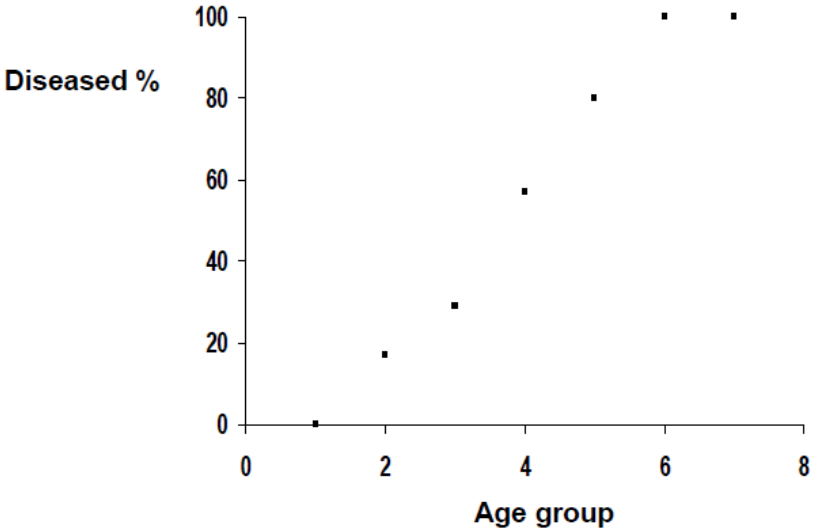


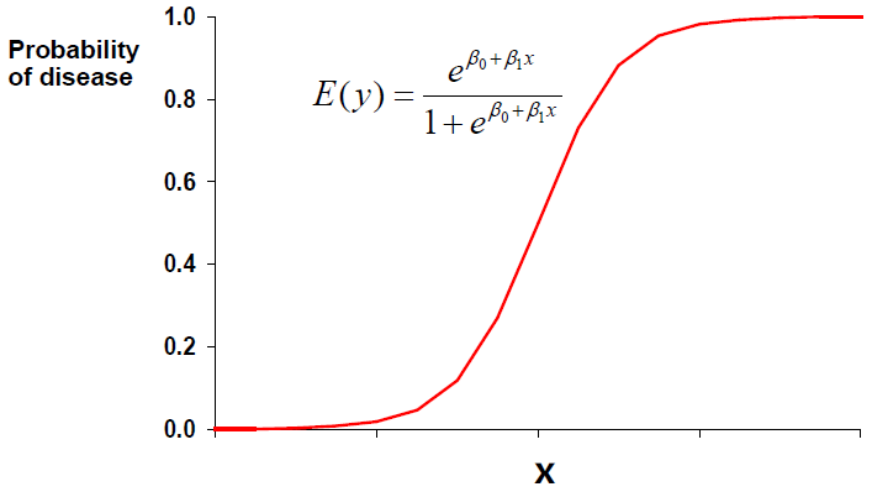




10.2 The Logistic Regression Model

- 시그모이드(Sigmoid) 함수 변환 예





10.2 The Logistic Regression Model

Odds(오즈)

- Odds(승산비): 클래스 '1'에 속할 오즈 → "클래스 '0'에 속하는 확률에 대한 클래스 '1'에 속하는 확률의 비"

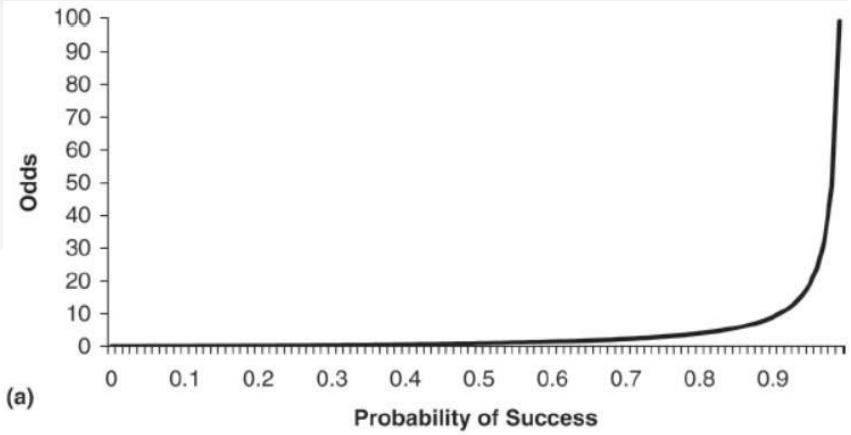
$$Odds(Y=1) = \frac{p}{1-p}$$

- 일반적으로 확률을 많이 사용하는 경마, 스포츠, 도박, 게임 등에서 오즈를 사용
- 사건이 발생하지 않을 확률 대비 사건이 발생할 확률
- 확률과 오즈의 관계

$$p = \frac{odds}{1+odds}$$

ex) $p = 0.5 \rightarrow odds = 1,$

$p = 0.7 \rightarrow odds = 2.3$



10.2 The Logistic Regression Model

로지스틱 모델 표준형

- 로지스틱 반응 함수

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q)}}$$

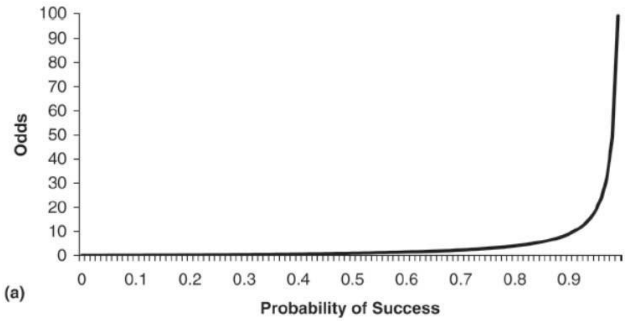
- 로지스틱 반응함수 식을 사용하여 오즈 식 정리
  - ✓ 예측변수와 오즈 사이의 승법적인(비례) 관계 설명 → 예측변수 X_j 가 한 단위 증가하면 다른 예측변수들이 모두 일정하다고 가정하면 오즈는 e^{β_j} 만큼 증가

$$Odds(Y = 1) = \frac{p}{1 - p} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q}$$

- 양변에 자연로그 취하여, 로지스틱 모델 표준형을 얻음
  - ✓ 로짓을 종속변수로 하여 q개의 예측변수에 대한 선형 함수로 모델링

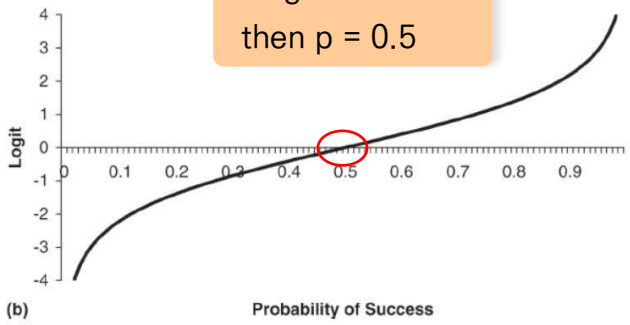
$$\textbf{logit} = \log(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

$$p = \frac{odds}{1 + odds} = \frac{e^{logit}}{1 + e^{logit}}$$



$$0 < Odds < \infty$$

if logit = 0
then p = 0.5



$$-\infty < logit < \infty$$

10.3 Example: Acceptance of Personal Loan

개인대출제안 수락

- 유니버셜(Universal) 은행은 채무가 있는 고객들을 개인대출 고객으로 전환하는 방법을 찾고자 함.
- 채무 고객들에게 펼친 캠페인에서 전환율: 9.6% (480/5000)
- 목적: 미래의 캠페인에서 대출제안에 긍정적으로 반응할 가능성이 높은 고객을 식별하는 모델 구축
- 데이터 셋: 5000명 (학습 데이터: 3000명)

ID	Age	Experience	Income	Family Size	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	Credit Card	
1	25	1	49	4	1.60	UG	0	0	1	0	0	0	0 or 1

✓ CCAvg: 월별 신용카드 평균사용액

10.3 Example: Acceptance of Personal Loan

Model with a Single Predictor

- 단일 예측변수를 갖는 간단한 로지스틱 회귀모델 구축
- 1개의 예측변수 X와 결과변수 Y의 관계를 직선으로 적합시킨 단순 선형회귀 모델과 개념적으로 유사
- 예측변수로 '수입(Income)'만 사용
- 확률의 관점에서 결과변수와 예측변수를 연관짓는 방정식

$$P(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

cf. 로지스틱 반응함수

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q)}}$$

$$Odds(Personal\ Loan = Yes \mid Income = x) = e^{\beta_0 + \beta_1 x}$$

- 모델로부터 추정된 계수: $\widehat{\beta_0} = -6.04892$ $\widehat{\beta_1} = 0.036$
- 적합시킨 모델

$$P(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{6.04892 - 0.036x}}$$

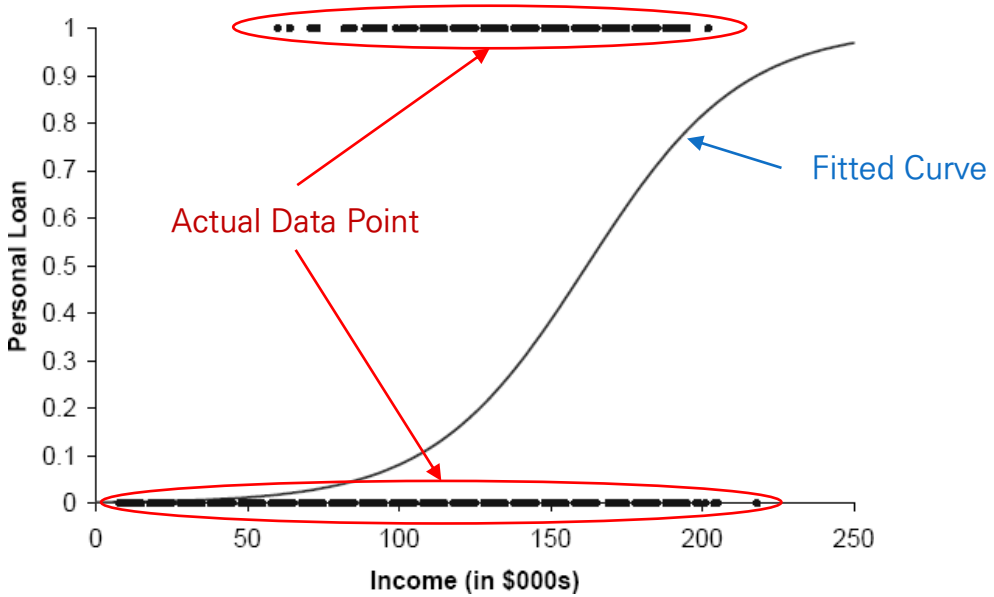
10.3 Example: Acceptance of Personal Loan

Model with a Single Predictor

- 적합시킨 모델

$$P(Personal\ Loan = Yes \mid Income = x) = \frac{1}{1 + e^{6.04892 - 0.036x}}$$

- 대출제안을 수락할 확률 추정: 새로운 고객이 대출제안을 수락할 고객인지, 거절할 고객인지 분류하기 위해 고객의 수입 정보를 위 식에 대입
- 고객이 대출제안을 수락할 확률이 cutoff보다 높다면 해당 고객은 대출제안을 수락할 고객으로 분류

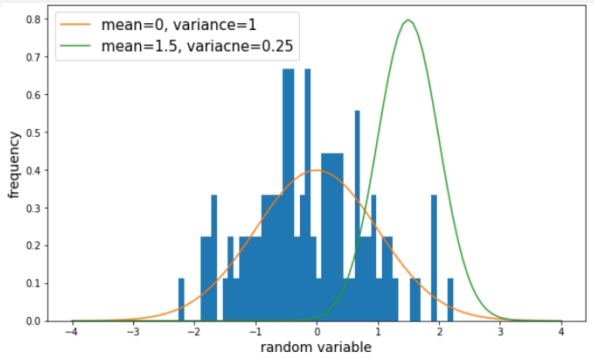


10.3 Example: Acceptance of Personal Loan

Estimating the Logistic Model from Data: Computing Parameter Estimates

- 패러미터 추정 값의 계산
- 로지스틱 회귀에서 종속변수 Y와 패러미터 β 의 관계는 비선형 → 패러미터 β 는 다중선형회귀와 같이 최소제곱법을 사용하여 추정되지 않음 → 최대가능도 방법(Maximum Likelihood Method)을 사용하여 추정
- 최대가능도 방법 (또는 최대우도법)(Maximum Likelihood Method)
  - ✓ 어떤 패러미터가 주어졌을 때, 원하는 값들이 나올 가능도를 최대로 만드는 패러미터를 선택하는 방법
- 회귀계수 추정치를 신뢰할 수 있는 경우 (일반적으로 선형회귀 알고리즘 보다 강건(robust)하지 못함)
  - ✓ 데이터에서 결과변수로 '0'과 '1'을 갖는 관측 값이 많을 때
  - ✓ 그 값들의 비율이 '0'과 '1' 둘 중 무엇과도 가깝지 않을 때
  - ✓ 로지스틱 회귀 모델에서 회귀계수의 개수가 표본의 크기에 비해 작을 때 (예) 표본의 10% 이하)

✓ **최대가능도 방법 (또는 최대우도법)** (Maximum Likelihood Method)
✓ 모수가 $\mu = 1.5, \sigma^2 = 0.25$인 정규분포보다 $\mu = 0, \sigma^2 = 1$ 인 정규분포에서 100개의 샘플이 추출되었을 **가능성이 높다**



10.3 Example: Acceptance of Personal Loan

Estimating the Logistic Model from Data: Computing Parameter Estimates

- 유니버설 은행 고객 3000명 대상 모델 적합
- 결과변수: Personal Loan (Yes = Success): 대출제안 수락 = 1, 거절 = 0

Data Preprocessing

- 'Education': 3개의 범주값('1', '2', '3') → 2개의 더미 변수 생성
- 5개의 범주형 변수 → 총 6개의 더미변수 생성(2 + 1 + 1 + 1 + 1)
- 수치형 예측변수: 6개
- 총 예측변수: 12개
- 학습 데이터: 60% / 검증 데이터: 40%

Age	Experience	Income	Family Size	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	Credit Card
25	1	49	4	1.60	UG	0	0	1	0	0	0

10.3 Example: Acceptance of Personal Loan

[실습] Table 10.2

Estimating the Logistic Model from Data: Computing Parameter Estimates

Estimated Model

- 양의 회귀계수: 해당 예측변수의 값이 클수록 제안을 수락할 확률이 높음
  - ✓ 'Education_Graduate', 'Education_Advanced/Professional', 'CD_Account': 대학원 교육 또는 전문 교육을 받은 것과 CD 계좌를 가진 것(더미변수에서 모두 '1'로 표기)이 대출 제안을 수락할 확률이 높음
- 음의 회귀계수: 해당 예측변수의 값이 클수록 제안을 수락할 확률이 낮음
  - ✓ 'Securities_Account', 'Credit Card': 증권계좌를 가진 것, 신용카드를 가진 것은 대출제안을 수락할 확률이 낮음

cf. $\text{logit} = \log(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$

[Model Fitting Result]

```
intercept  -12.493436061176814
              Age  Experience    Income    Family     CCAvg  Mortgage
coeff -0.037685    0.039202  0.058844  0.612251  0.240489  0.001012

          Securities_Account  CD_Account    Online  CreditCard  ₩
coeff                -1.01428    3.649097 -0.678306   -0.958283

          Education_Graduate  Education_Advanced/Professional
coeff               4.202148                         4.355761

AIC -709.1524769205962
```

$$logit(Personal\ Loan = Yes)$$
$$= -12.493 - 0.0377\ Age + 0.0392\ Experience$$
$$+ \cdots + 4.356\ Education\_Advanced/Professional$$

10.3 Example: Acceptance of Personal Loan

Interpreting Results in Term of Odds (for a Profiling Goal)

- 오즈 관점에서의 결과 해석
- 로지스틱 모델은 서로 다른 예측변수들의 역할에 대해 유용한 정보를 줄 수 있음

  ex) 수입(Income)이 한 단위 증가하는 것이 대출 제안을 수락할 확률에 어떠한 영향을 미치는지 알고 싶다.
- 오즈 식은 다음과 같이 표현

  $Odds = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_q x_q}$
- 고객의 수입에 따라 개인대출제안의 수락여부를 모델링한 경우

  $Odds(Personal\ Loan = Yes \mid Income) = e^{\beta_0+\beta_1 Income}$
- Base-case odds(기준 오즈): 수입이 0인 고객이 대출제안을 수락할 오즈

  $e^{-6.04892+(0.036)(0)} = 0.00236$
- 수입이 $100K인 고객이 대출제안을 수락할 오즈는 수입이 0일 때 보다 36.6배($e^{(0.036)(100)} = 36.6$) 증가함
- 이 고객이 대출제안을 수락할 오즈

  $e^{-6.04892+(0.036)(100)} = 0.086$

10.3 Example: Acceptance of Personal Loan

Interpreting Results in Term of Odds (for a Profiling Goal)

- 수입: x_1, 다른 예측변수: $(x_2, \cdots, x_{12})$ (일정한 값으로 고정)
- 이때 수입이 x_1에서 x_{1+1} 로 한 단위 증가하는 경우
- 오즈 비

$$\frac{odds(x_1+1, x_2, \cdots, x_{12})}{odds(x_1, x_2, \cdots, x_{12})} = \frac{e^{\beta_0+\beta_1(x_1+1)+\beta_2 x_2+\cdots+\beta_{12}x_{12}}}{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_{12}x_{12}}} = e^{\beta_1}$$

cf. $Odds = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_q x_q}$

- x_1 이 한 단위 증가할 때, 고객이 대출제안을 수락할 오즈가 e^{β_1} 만큼 증가함
- e^{β_1} : x_1 이외의 모든 예측변수가 고정되어 있고 x_1 이 1 단위 증가할 때, 클래스 '1'에 속할 오즈가 증가하는 곱셈 요인
- $\beta_1 > 0 \rightarrow$ 오즈의 증가 $\quad\quad$ $\beta_1 < 0 \rightarrow$ 오즈의 감소

10.3 Example: Acceptance of Personal Loan

Interpreting Results in Term of Odds (for a Profiling Goal)

- **더미변수에 대한 해석**은 기술적으로는 동일하지만 실질적으로는 다른 의미
- ex) CD_Account의 회귀계수 = 3.649097
- 기준 그룹: CD 계좌를 보유하고 있지 않은 고객
- $e^{3.649097} = 38.4$ 의 해석
  - ✓ 다른 요인들은 모두 일정한 것으로 가정할 때, CD 계좌를 가진 고객이 가지지 않은 고객에 비하여 대출제안을 수락할 오즈
  - ✓ CD 계좌를 가진 고객이 가지지 않은 고객에 비하여 대출제안을 수락할 경향이 더 높음
- 확률과 대조적으로 모델의 결과를 오즈로 보고할 때의 장점: x_1 의 어떤 값에 대해서도 위와 같은 해석이 가능
- 확률의 경우, x_1 이 더미변수가 아니라면, x_1의 한 단위 증가가 확률에 미치는 효과를 위와 같이 설명할 수 없음
  - ✓ x_1이 3에서 4로 증가할 때 클래스 '1'에 속하는 확률 p에 미치는 효과와 30에서 31로 증가할 때 확률 p에 미치는 효과는 다름
  - ✓ 다른 예측변수가 모두 일정하다고 했을 때, 특정 예측변수의 한 단위 증가에 의한 확률 p의 변화는 일정하지 않음 → 확률의 변화는 예측변수의 특정 값에 따라 다름 → 오직 특정 레코드에 대해서만 확률이 언급될 수 있음

10.4 Evaluating Classification Performance

[실습] Table 10.3

Confusion Matrix

- Confusion matrix 작성: 추정된 로지스틱 회귀식으로부터 검증 데이터의 각 레코드에 대하여 클래스에 속할 확률(성향)을 예측 → cutoff 값을 사용하여 클래스 결정

$$logit(Personal\ Loan = Yes)$$
$$= -12.493 - 0.0377\ Age + 0.0392\ Experience$$
$$+ \cdots + 4.356\ Education\_Advanced/Professional$$

Logit 값 구한 후 확률 p 계산 →

$$p = \frac{e^{logit}}{1 + e^{logit}}$$

- [오른쪽 예] 검증 데이터: 2000 (거절: 0, 수용: 1)
- if cutoff 값 = 0.5
- 고객 2764: $p(1) = 0.024062 < 0.5$ → 클래스 '0' 분류
- 고객 932, 2721: $p(1) > 0.5$ → 클래스 '1' 분류(오분류: 932)
- 고객 702: $p(1) < 0.5$ → 클래스 '0' 분류(오분류: 702)

검증 데이터의 첫 4개의 레코드의 확률

	actual	p(0)	p(1)	predicted
2764	0	0.975938	0.024062	0
932	0	0.331932	0.668068	1
2721	1	0.031478	0.968522	1
702	1	0.985949	0.014051	0

Misclassified

10.4 Evaluating Classification Performance

[실습] Table 10.4, Figure 10.3

Confusion Matrix

- 실제로 '0'(거절)인 클래스 → 분류 성능 좋음. 오분류율: 약 1%
- 실제로 '1'(수용)인 클래스 → 오분류율: 약 1/3

Cutoff = 0.5

학습 데이터

Confusion Matrix (Accuracy 0.9607)

Actual \ Prediction	0	1
0	2685	28
1	90	197

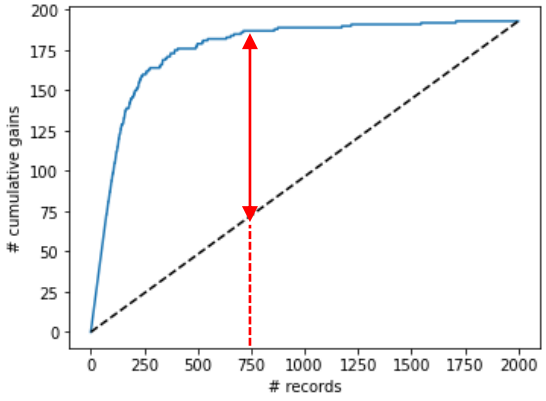
검증 데이터

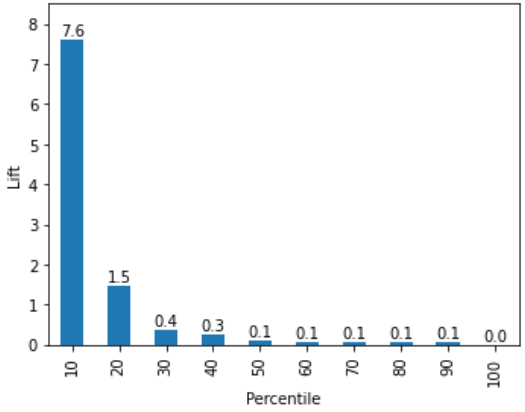
Confusion Matrix (Accuracy 0.9595)

Actual \ Prediction	0	1
0	1791	16
1	65	128

Cumulative Gain and Decile Lift Chart

- Cumulative Gain: baseline에 비해 왼쪽과 위쪽으로 치우친 좋은 성능을 보임
  - ✓ Baseline 위로 'lift'된 양: 이 모델에 의해 순위를 매긴 총 레코드의 x개(예를 들어 750)를 택하면, 모델을 사용하지 않은 경우보다 실제로 '1'(수용)인 클래스의 레코드를 추가로 얻을 것으로 기대되는 양
- 10분위 차트: 모델에서 '1'(수용)으로 분류될 확률이 높은 상위 10% 레코드를 선택한 경우, 임의로 선택한 경우보다 대출제안을 수락할 고객이 7.6배 높음





10.4 Evaluating Classification Performance

Variable Selection

- 다음 단계: 대체 모델 탐색
  - ✓ 예측변수의 개수를 줄이면서 더 단순한 모델 탐색
  - ✓ 예측변수들 간의 상호작용을 고려하고 그로부터 새롭게 파생되는 변수들을 고려함으로써 더 복잡한 모델 구축
  - ✓ ex) 수입과 가족 규모 사이에 교호작용(interactive effect)이 있다고 가정하면, '수입 X 가족'의 형태로 교호작용 항 추가
- 자동화된 변수 선택 방법 사용
  - ✓ 단계적 선태(Stepwise Selection), 전진 선택(Forward Selection), 후진제거(Backward Elimination) (선형회귀에서 사용한 방법과 동일)
  - ✓ AIC나 다른 측도들의 값을 최소화: 예측변수의 개수를 고려함으로써 학습 데이터에 적합하는 것에 대한 벌점(penalty)를 부여하는 방법 사용
- 규제(regularization) 방법 사용: L1 과 L2 벌점(penalty) 이용
  - ✓ 라쏘 회귀(Lasso regression): L1 penalty
  - ✓ 릿지 회귀(Ridge regression): L2 penalty

10.5 Logistic Regression for Multi-class Classification

- 각 관측 값에 대해 m개의 클래스에 속할 m개의 확률 가정
- m개의 확률에 대한 합이 1이기 때문에 단지 m-1개의 확률만 추정
- Ordinal Classes: 클래스의 순서가 의미있는 경우 (ex) 주식 추천: 매수(buy), 보유(hold), 매도(sell))
- Nominal Classes: 클래스의 순서가 의미없는 경우 (ex) 다수브랜드의 시리얼 중 선택)

Ordinal Classes(순서형 클래스)

- 순서형 클래스: 의미있는 순서를 가지고 있는 클래스. 간단한 규칙으로 클래스에 의미있는 방법으로 번호 부여 가능 ex) 주식 추천: 매수(buy), 보유(hold), 매도(sell)
- 클래스의 개수에 따른 방법
  - ✓ 클래스의 수가 많을 때(일반적으로 $m > 5$): 종속변수를 연속형으로 취급하여 다중선형회귀 수행
  - ✓ $m = 2$ 일 때: 로지스틱 회귀 사용
  - ✓ $3 \leq m \leq 5$ 일 때: 로지스틱 회귀의 확장 필요 → **비례 오즈(Proportional Odds) 또는 누적 로짓(Cumulative Logit) 방법** 사용

10.5 Logistic Regression for Multi-class Classification

Ordinal Classes(순서형 클래스)

- ex) 주식 추천: $m = 3$ 클래스 가정 (1=매수(buy), 2=보유(hold), 3=매도(sell))

$$P(buy) = P(Y \leq 1) \qquad P(buy\ or\ hold) = P(Y \leq 2)$$

1. 클래스 소속도에 대한 3개의 비누적 확률은 2개의 누적 확률로부터 획득

$$P(Y = 1) = P(Y \leq 1)$$

$$P(Y = 2) = P(Y \leq 2) - P(Y \leq 1)$$

$$P(Y = 3) = 1 - P(Y \leq 2)$$

2. 각 logit을 예측변수의 함수로서 모델링

$$\text{logit}(buy) = \log\frac{P(Y \leq 1)}{1 - P(Y \leq 1)} \qquad \text{logit}(buy\ or\ hold) = \log\frac{P(Y \leq 2)}{1 - P(Y \leq 2)}$$

cf. $\textbf{logit} = \log(odds) = \log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$

10.5 Logistic Regression for Multi-class Classification

Ordinal Classes(순서형 클래스)

2. 각 logit을 예측변수의 함수로서 모델링

$$\text{logit}(buy) = \log\frac{P(Y \le 1)}{1 - P(Y \le 1)} \qquad \text{logit}(buy\ or\ hold) = \log\frac{P(Y \le 2)}{1 - P(Y \le 2)}$$

3. 각 logit은 예측변수들의 선형함수로 모델링. 주식추천이 단일 예측변수 x 를 갖는다면, 아래의 식을 구할 수 있음

$$\text{logit}(buy) = \alpha_0 + \beta_1 x \qquad \text{logit}(buy\ or\ hold) = \beta_0 + \beta_1 x$$

$$\text{cf. } \textbf{logit} = \log(odds) = \log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$$

- 두 식의 기울기(β_1)는 같지만, 절편(α_0 , β_0)은 다름('buy' 또는 'buy or hold'에 따라 달라짐)
- 회귀계수 $\alpha_0, \beta_0, \beta_1$ 이 추정되면, 확률의 관점에서 로짓 방정식을 다시 표현함으로써 클래스에 속할 확률을 구할 수 있음

10.5 Logistic Regression for Multi-class Classification

Ordinal Classes(순서형 클래스)

cf. $\textbf{logit} = \log(odds) = \log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$

3. 각 logit은 예측변수들의 선형함수로 모델링. 주식추천이 단일 예측변수 x 를 갖는다면, 아래의 식을 구할 수 있음

$\text{logit}(buy) = \alpha_0 + \beta_1 x$ $\quad\quad$ $\text{logit}(buy\ or\ hold) = \beta_0 + \beta_1 x$

4. 이 예제에서 3개의 클래스에 대해서는 다음과 같은 식을 얻을 수 있음

$$P(Y=1) = P(Y \le 1) = \frac{1}{1+e^{-(a_0+b_1 x)}}$$

$$P(Y=2) = P(Y \le 2) - P(Y \le 1) = \frac{1}{1+e^{-(b_0+b_1 x)}} - \frac{1}{1+e^{-(a_0+b_1 x)}}$$

$$P(Y=3) = 1 - P(Y \le 2) = 1 - \frac{1}{1+e^{-(b_0+b_1 x)}}$$

(a_0, b_0, b_1 : 학습 데이터로부터 얻어진 추정 값)

5. 관측 값을 3개의 클래스 중 하나로 분류

- 주식의 확률 추정값이 $P(Y=1) = 0.2, P(Y=2) = 0.3, P(Y=3) = 0.5$ 와 같다면, 이 주식은 '매도' 추천으로 분류

10.5 Logistic Regression for Multi-class Classification

Nominal Classes(명목형 클래스)

- 명목형 클래스 확인 방법: A, B, C, … 로 태그를 만들고, 클래스에 문자를 할당하는 것이 중요하지 않은 경우
- ex) 시리얼 제품의 여러 가지 브랜드 사이의 선택
- 가정: 소비자가 선택할 수 있는 시리얼 브랜드의 개수 $m = 3$ (각 소비자가 1개의 브랜드 선택 가정)
- 시리얼의 가격 x 를 알 때, 어떤 시리얼이 선택될 지 예측
- $P(Y = A), P(Y = B), P(Y = C)$ 의 확률 추정 가능
- 앞의 예와 마찬가지로, 3개의 클래스 중 하나를 기준 클래스로 사용. (여기서는 C 사용)

1. 예측변수와 선형인 $m - 1$ 개의 유사 로짓(pseudo logit) 식 도출

$$\text{logit}(A) = \log\frac{P(Y = A)}{P(Y = C)} = \alpha_0 + \alpha_1 x$$

$$\text{logit}(B) = \log\frac{P(Y = B)}{P(Y = C)} = \beta_0 + \beta_1 x$$

10.5 Logistic Regression for Multi-class Classification

Nominal Classes(명목형 클래스)

1. 예측변수와 선형인 $m-1$ 개의 유사 로짓(pseudo logit) 식 도출

$$\text{logit}(A) = \log\frac{P(Y=A)}{P(Y=C)} = \alpha_0 + \alpha_1 x$$

$$\text{logit}(B) = \log\frac{P(Y=B)}{P(Y=C)} = \beta_0 + \beta_1 x$$

2. 위 식으로부터 아래 식을 구할 수 있음

$$P(Y=A) = P(Y=C) \times e^{\alpha_0+\alpha_1 x}$$

$$P(Y=B) = P(Y=C) \times e^{\beta_0+\beta_1 x}$$

3. $P(Y=A)+P(Y=B)+P(Y=C)=1$ 을 만족하므로,

$$P(Y=C) \times e^{\alpha_0+\alpha_1 x} + P(Y=C) \times e^{\beta_0+\beta_1 x} + P(Y=C) = 1$$

$$P(Y=C) = \frac{1}{1+e^{\alpha_0+\alpha_1 x}+e^{\beta_0+\beta_1 x}}$$

4. 학습 데이터를 이용하여 4개의 회귀계수를 추정할 때, 각 클래스에 속할 확률은 다음과 같이 추정됨

$$P(Y=A) = \frac{e^{a_0+a_1 x}}{1+e^{a_0+a_1 x}+e^{b_0+b_1 x}}$$

$$P(Y=B) = \frac{e^{b_0+b_1 x}}{1+e^{a_0+a_1 x}+e^{b_0+b_1 x}}$$

$$P(Y=C) = 1 - P(Y=A) - P(Y=B)$$

10.5 Logistic Regression for Multi-class Classification

Comparing Ordinal and Nominal Models

- 결과변수가 순서형 또는 명목형으로 되어 있는지에 따라 추정된 로지스틱 모델이 다름
- ex) 자동차 사고 데이터(from US Bureau of Transportation Statistics)
- 알코올, 시간대, 도로 조건 등의 예측변수에 기초하여, 어떤 사고가 부상 또는 사망을 유발하는지 예측
- 사고의 심각성: 상해 없음(No injury = 0), 부상(Nonfatal injuries = 1), 사망(Fatalities = 2)

- 사고의 심각성을 순서형과 명목형으로 간주하여 실험
- 2개의 명목형 예측변수 사용: 알코올, 날씨 (No: 사고에 연관되어 있지 않음, Yes:사고와 연관되어 있음)
- 기준 클래스(심각성 수준): 심각성 = 2

10.5 Logistic Regression for Multi-class Classification

Comparing Ordinal and Nominal Models

- 2개의 예측변수 사용: 알코올, 날씨
- 명목형 로지스틱 모델
  - ✓ 절편: 3개 추정
  - ✓ 회귀계수: 결과변수(심각성)의 수준 별 3개의 서로 다른 회귀계수 집합 추정

$$\text{logit}(0) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$$

$$\text{logit}(1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\text{logit}(2) = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2$$

- 순서형 로지스틱 모델
  - ✓ 절편: 2개 추정 (심각성: 상해 없음(0), 부상(1))
  - ✓ 회귀계수: 2개 추정(각 예측변수에 대해서 1개)

$$\text{logit}(0) = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\text{logit}(1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

cf. $p = \frac{e^{logit}}{1+e^{logit}}$

Nominal

```
predictors ['ALCHL_I', 'WEATHER_R']

intercept [-0.09100315  0.9036454  -0.81264225]
coef [[ 0.51606685  0.3391015 ]
 [ 0.14900396  0.09543369]
 [-0.66507082 -0.43453518]]
classes [0 1 2]
```

	actual	predicted	P(0)	P(1)	P(2)
0	1	1	0.490649	0.498989	0.010362
1	0	0	0.553461	0.441147	0.005392
2	0	0	0.553461	0.441147	0.005392
3	0	1	0.490649	0.498989	0.010362
4	0	1	0.394192	0.578684	0.027124

Ordinal

```
theta [-1.06916285  2.77444326]
coef [-0.40112008 -0.25174207]
classes [0 1 2]
```

	actual	predicted	P(0)	P(1)	P(2)
0	1.000614	1	0.496856	0.482272	0.020529
1	-0.000829	0	0.559252	0.422993	0.016207
2	-0.000384	0	0.560104	0.424920	0.014911
3	-0.000377	1	0.495261	0.482900	0.023268
4	-0.001671	1	0.397167	0.572164	0.030877

- 결과적으로 결과변수를 순서형으로 또는 명목형으로 간주할 것이냐의 문제는 검증 데이터에 대한 예측 성능의 평가에 달려 있음

10.6 Example of Complete Analysis : Predicting Delayed Flights

- 공항, 항공사, 항공 당국 등은 비행기 연착에 대한 예측을 통해 연착이 예상되는 비행기에 대해 사전 조치를 취할 수 있음
- 목적: 새로운 항공편이 지연(15분 이상 늦게 도착하는 것)될지 아닐지 예측
- 2004년 1월 워싱턴 D.C. 지역에서 뉴욕시 지역으로 가는 모든 운항(from 교통통계국)
- 전체 2,201편 운항 / 지연된 운항: 19.5%
- 결과변수: delayed(지연) or not (1 = delayed / 0 = on time)

변수명	변수명	변수 내역
Day of week	요일	1=월요일, 2=화요일, …, 7=일요일
Departure time	출발시간	오전 6시와 오후 10시 사이를 16 구간으로 나눔.
Origin	출발공항	3개의 공항코드: DCA(레이건 내셔널), IAD(덜레스), BWI(볼티모어-워싱턴 국제)
Destination	도착공항	3개의 공항코드: JFK(케네디), LGA(라구아디아), EWR(뉴어크)
Carrier	항공회사	8개의 항공사 코드: CO(컨티넨탈), DH(아틀란틱 코우스트), DL(델타), MQ(아메리칸 이글), OH(컴에어), RU(컨티넨탈 익스프레스), UA(유나이티드), US(유에스 에어웨이즈)
Weather	날씨	악천후로 연기된 경우 '1'로 표기

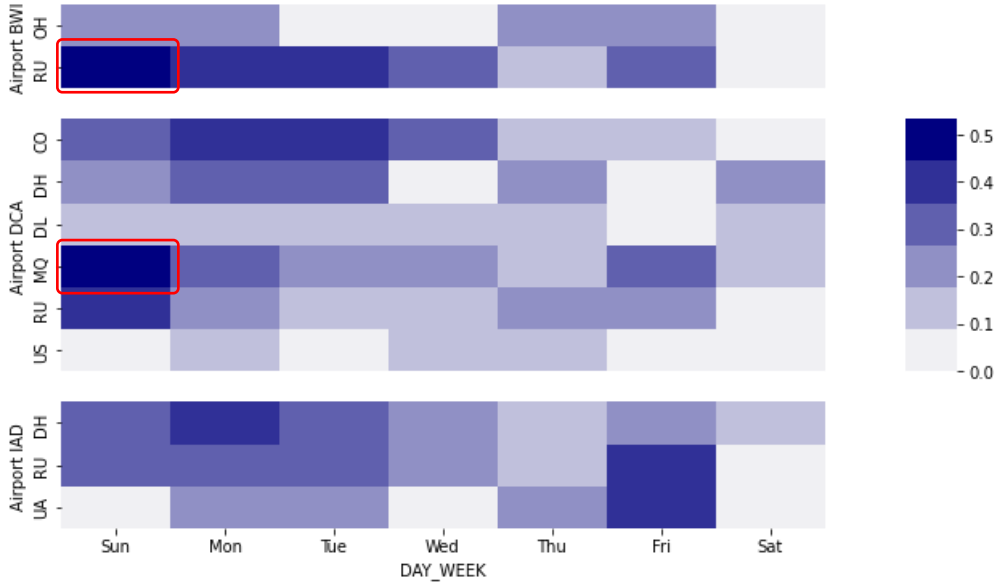
10.6 Example of Complete Analysis : Predicting Delayed Flights

[실습] Figure 10.5

Flight Status	Carrier	Day of Week	Departure time	Destination	Origin	Weather
ontime	OH	4	1455	JFK	BWI	0
ontime	DH	4	1640	JFK	DCA	0
ontime	DH	4	1245	LGA	IAD	0
ontime	DH	4	1709	LGA	IAD	0
ontime	DH	4	1035	LGA	IAD	0

연착 항공편 비율 Heat Map (Grouped by Day of week, Origin, carrier)

- 높은 연착률 조합
  - ✓ BWI에서 출발하여 RU 항공사로 일요일 출발
  - ✓ DCA에서 출발하여 MQ 항공사로 일요일 출발

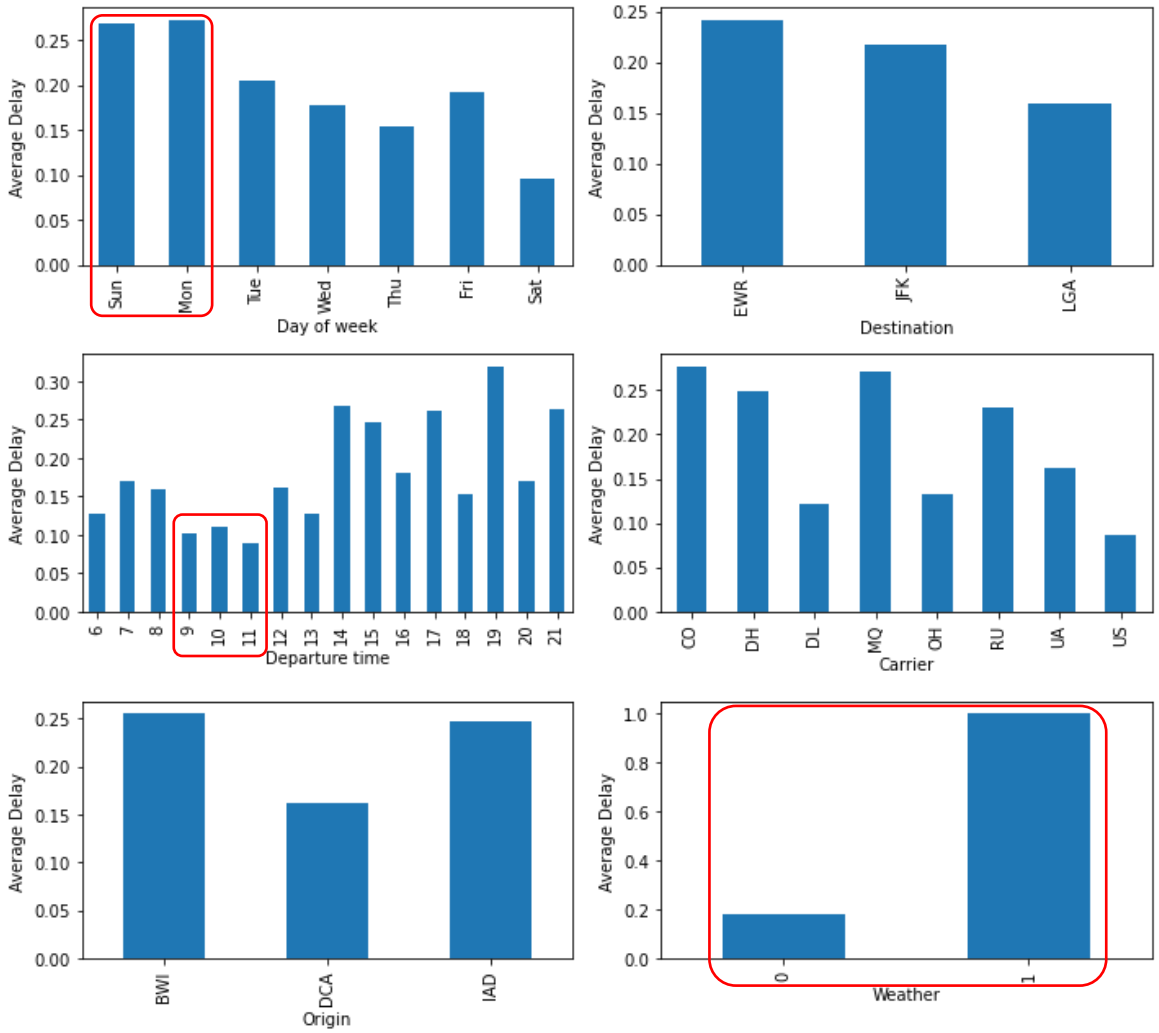


10.6 Example of Complete Analysis : Predicting Delayed Flights

[실습] Figure 10.4

- Day of week
  - ✓ 일요일과 월요일이 연착 비율 높음
- Departure time
  - ✓ 오전 9시 부터 11시 사이에 출발하는 항공편의 연착 비율 낮음
- Weather
  - ✓ '0' 과 '1'에 따라 연착여부가 뚜렷하게 구분

6개 예측변수 별 연착 항공편 비율 (출발시간은 1시간으로 구분)



10.6 Example of Complete Analysis : Predicting Delayed Flights

분석의 주 목적

- 예측변수의 정보를 기반으로 새로운 항공편의 연착 여부 분류
- 연착될 가능성이 가장 높거나/낮은(ranking) 항공편의 비율 찾음
- 어떤 요인이 연착과 관련이 있는지(이 sample 뿐만 아니라 해당 노선의 전체 항공편에 대해서) 찾아내고 이들 요인을 정량화

Flight Status	Carrier	Day of Week	Departure time	Destination	Origin	Weather
ontime	OH	4	1455	JFK	BWI	0
ontime	DH	4	1640	JFK	DCA	0

Data Preprocessing

- 이진형 결과변수 생성: isDelay (if Flight Status = delayed → '1' , = ontime → '0')
- Day of week(요일): 범주형 변수로 변환
- Departure time(출발시간): 오전 6시에서 오후 10시까지 1시간 단위로 범주화
- 범주형 변수에 대해 기준 범주 설정: 출발공항–BWI, 도착공항–EWR, 항공사–CO(컨티넨탈)
- 총 33개의 예측변수
- 학습용 데이터(60%), 검증용 데이터(40%)

10.6 Example of Complete Analysis : Predicting Delayed Flights

[실습] Table 10.7

Model Training

변수명	변수명	변수 내역	범주 개수	예측변수(총 33개)	기준 범주
Day of week	요일	1=월요일, 2=화요일, …, 7=일요일	7	6	월요일
Departure time	출발시간	오전 6시와 오후 10시 사이 16 구간	16	15	06-07
Origin	출발공항	3개: BWI, DCA, IAD	3	2	BWI
Destination	도착공항	3개: EWR, JFK, LGA	3	2	EWR
Carrier	항공회사	8개: CO, DH, DL, MQ, OH, RU, UA, US	8	7	CO
Weather	날씨	악천후로 연기된 경우 '1'로 표기	2	1	

```
intercept  -1.2190996255195397
       Weather  DAY_WEEK_2  DAY_WEEK_3  DAY_WEEK_4  DAY_WEEK_5  DAY_WEEK_6  DAY_WEEK_7  ₩
coeff    9.325      -0.598      -0.705      -0.799      -0.296      -1.129      -0.135

       CRS_DEP_TIME_7  CRS_DEP_TIME_8  CRS_DEP_TIME_9  CRS_DEP_TIME_10  CRS_DEP_TIME_11  ₩
coeff           0.631           0.382          -0.365            0.337            0.078

       CRS_DEP_TIME_12  CRS_DEP_TIME_13  CRS_DEP_TIME_14  CRS_DEP_TIME_15  CRS_DEP_TIME_16
coeff            0.399            0.175            0.202            1.265            0.628

       CRS_DEP_TIME_17  CRS_DEP_TIME_18  CRS_DEP_TIME_19  CRS_DEP_TIME_20  CRS_DEP_TIME_21
coeff            1.093            0.285            1.655            1.023            1.077

       ORIGIN_DCA  ORIGIN_IAD  DEST_JFK  DEST_LGA  CARRIER_DH  CARRIER_DL  CARRIER_MQ  ₩
coeff      -0.01      -0.134    -0.524    -0.546       0.352      -0.685       0.743

       CARRIER_OH  CARRIER_RU  CARRIER_UA  CARRIER_US
coeff      -0.711      -0.194       0.315      -0.971

AIC 1004.5346225948085
```

10.6 Example of Complete Analysis : Predicting Delayed Flights

[실습] Table 10.7

Model Interpretation

- 도착공항 (기준 범주: EWR)
  - ✓ 도착공항이 JFK(DEST_JFK)인 경우에 대한 회귀계수 = −0.524
  - ✓ $e^{-0.524} = 0.59$: 다른 요인이 모두 일정할 때, EWR로 도착하는 항공편(기준 항공편)이 연착될 오즈에 대해 JFK로 도착하는 항공편이 연착될 오즈
  - ✓ 값이 1보다 작기 때문에 JFK로 도착하는 항공편이 EWR로 도착하는 항공편 보다 연착될 가능성이 작다

```
intercept  -1.2190996255195397
          Weather  DAY_WEEK_2  DAY_WEEK_3  DAY_WEEK_4  DAY_WEEK_5  DAY_WEEK_6  DAY_WEEK_7  ₩
coeff       9.325      -0.598      -0.705      -0.799      -0.296      -1.129      -0.135

          CRS_DEP_TIME_7  CRS_DEP_TIME_8  CRS_DEP_TIME_9  CRS_DEP_TIME_10  CRS_DEP_TIME_11  ₩
coeff              0.631           0.382          -0.365            0.337            0.078

          CRS_DEP_TIME_12  CRS_DEP_TIME_13  CRS_DEP_TIME_14  CRS_DEP_TIME_15  CRS_DEP_TIME_16
coeff               0.399            0.175            0.202            1.265            0.628

          CRS_DEP_TIME_17  CRS_DEP_TIME_18  CRS_DEP_TIME_19  CRS_DEP_TIME_20  CRS_DEP_TIME_21
coeff               1.093            0.285            1.655            1.023            1.077

          ORIGIN_DCA  ORIGIN_IAD  DEST_JFK  DEST_LGA  CARRIER_DH  CARRIER_DL  CARRIER_MQ  ₩
coeff         -0.01      -0.134    -0.524    -0.546       0.352      -0.685       0.743

          CARRIER_OH  CARRIER_RU  CARRIER_UA  CARRIER_US
coeff         -0.711      -0.194       0.315      -0.971

AIC 1004.5346225948085
```

- 항공사 (기준 범주: CO)
  - ✓ US: 가장 큰 음의 회귀계수(−0.971)를 가짐 → 가장 낮은 연착될 오즈
  - ✓ MQ: 가장 큰 연착될 오즈를 가짐
- 요일 (기준 범주: 월요일)
  - ✓ 월요일(DAY_WEEK_1): 다른 모든 요일이 음의 회귀계수를 가짐 → 가장 높은 연착될 오즈
  - ✓ 토요일(DAY_WEEK_6): 가장 큰 음의 회귀계수를 가짐 → 가장 낮은 연착될 오즈를 가짐
- 시간 (기준 범주: 오전 6-7시)
  - ✓ 오후 7-8시 (CRS_DEP_TIME_19): 기준 범주와 가장 큰 차이를 보임

10.6 Example of Complete Analysis : Predicting Delayed Flights

[실습] Figure 10.6

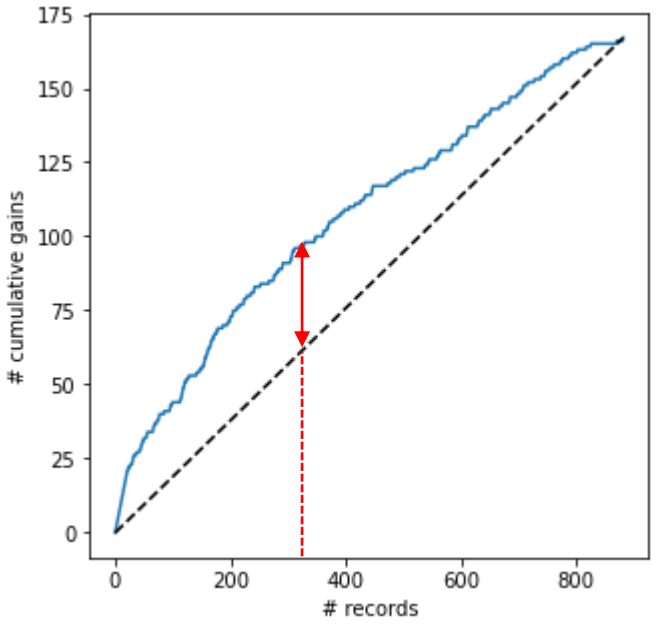
Model Performance

- Confusion Matrix
  - ✓ 연착된 항공편 보다 정시에 도착한 항공편을 더 정확하게 분류하고 있음
  - ✓ 연착 항공편 오분류율 = 27/(140+27) = 16.17%
  - ✓ 정시 도착 항공편 오분류율 = 9/(705+9) = 1.26%
- Cumulative Gain Chart
  - ✓ 연착을 예측하는 모델이 평균 연착률을 예측하는 단순한 모델에 비해 우수함(기준선 보다 우수한 성과)

검증 데이터

Confusion Matrix (Accuracy 0.8309)

	Prediction	
Actual	ontime	delayed
ontime	705	9
delayed	140	27



10.6 Example of Complete Analysis : Predicting Delayed Flights

[실습] Figure 10.5

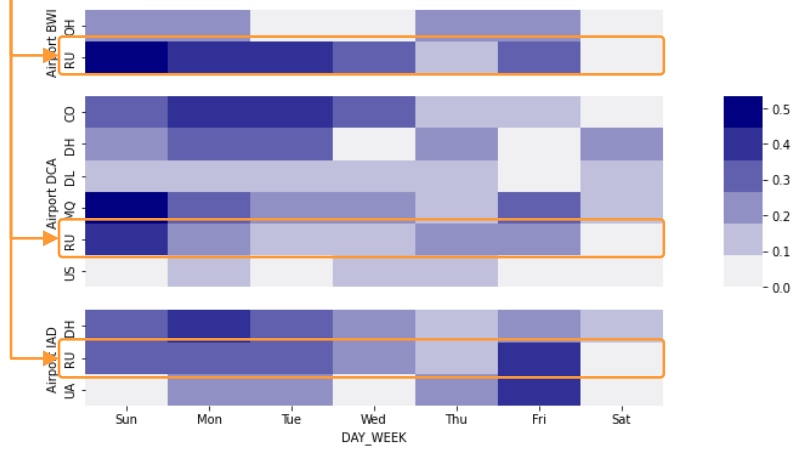
Variable Selection

- 출발공항 더미변수 제외
  - ✓ 대부분의 항공사는 한 공항(DCA)에서 출발함
  - ✓ 3개 공항 모두에서 출발하는 항공사(RU)는 출발공항이 어디인지와 관계없이 연착률은 비슷하게 나타남
- 도착공항 더미변수 제외
  - ✓ 실질적으로 모든 항공사가 모든 공항으로 비행하지 않음
  - ✓ 추정된 모델이 존재하지 않는 항공기와 도착공항의 조합에 대해 예측을 한다면 타당하지 않음

항공사와 출발공항 별 항공편 개수

	BWI	DCA	IAD	Total
CO		94		94
DH		27	524	551
DL		388		388
MQ		295		295
OH	30			30
RU	115	162	131	408
UA			31	31
US		404		404
Total	145	1370	686	2201

연착 항공편 비율 Heat Map
(Grouped by Day of week, Origin, carrier)

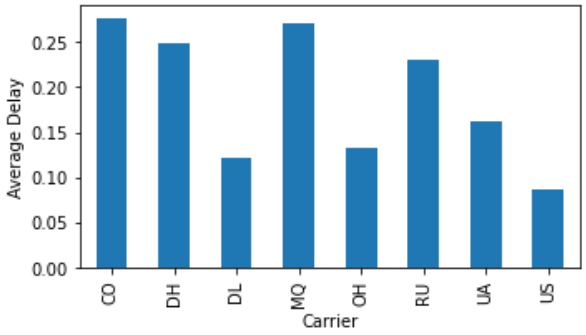


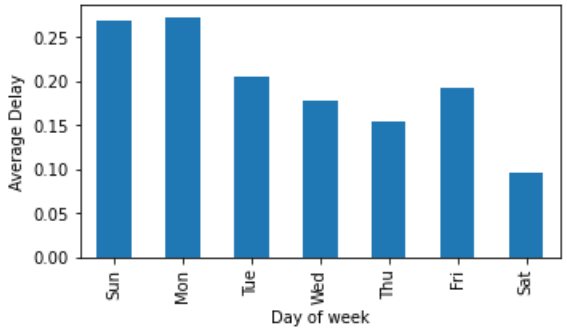
10.6 Example of Complete Analysis : Predicting Delayed Flights

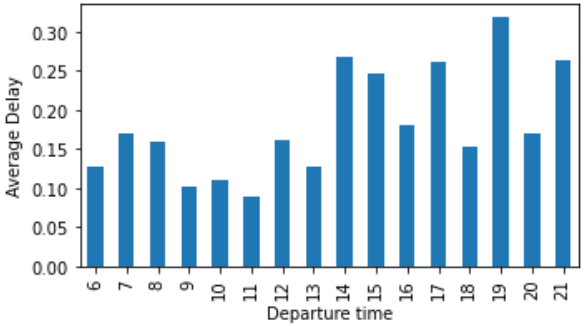
[실습] Figure 10.5

Variable Selection

- 항공사: {CO,MQ, DH, RU} = 1 / others = 0
- 요일: {월요일, 일요일} = 1 / others = 0
- 출발시간: {6, 7, 8, 9} = MORNING, {10, 11, 12, 13} = NOON, {14, 15, 16, 17, 18} = AFTER2P, {19, 20} = EVENING
- 자동화된 변수 선택 방법 사용: L1 Penalty를 갖는 규제(Regularization) → 7개 예측변수







10.6 Example of Complete Analysis : Predicting Delayed Flights

[실습] Table 10.9

Variable Selection

- 축소된 모델이 훨씬 적은 정보량을 이용함에도 불구하고 분류 정확도와 향상 측면에서 원 모델과 차이가 없음
- 예제의 결론: 항공사, 요일, 출발시간(4개), 악천후의 7개 예측변수 사용
- 날씨 변수: 미리 알 수 없음.
  - ✓ 분석 목적이 특정한 항공편이 지연될 것인지 미리 예측하는 것이라면 제거되어야 함
  - ✓ 분석 목적이 연착되지 않은 비행과 비교하여 연착된 경우를 프로파일링하는 경우라면, 날씨 변수를 일정하게 하여 다른 요인의 영향을 평가할 수 있도록 포함 가능(악천후가 있는 날과 없는 날 비교)
- 워싱턴 D.C.에서 뉴욕까지 정시에 도착할 가능성이 가장 높은 항공편: 화요일부터 토요일까지 오전 9시 부터 정오 사이에 운항하는 DL(델타), OH(컴에어), UA(유나이티드), US(유에스 에어웨이즈) 항공

적은 예측변수들로 적합한 로지스틱 회귀 모델 결과

```
regularization [2.7825594]
intercept  -2.2873909162550 59
       Sun_Mon  Weather CARRIER_CO_MQ_DH_RU  MORNING     NOON   AFTER2P   EVENING
coeff 0.577932 4.977922              1.2988 -0.583262 -0.665839 -0.055142 0.560885
AIC 934.6153607819033
Confusion Matrix (Accuracy 0.8343)

         Prediction
 Actual  ontime delayed
 ontime     711       3
delayed     143      24
```