4.3 - Logistic Regression

분류는 수치화할 수 없는 경우가 많다.
선형회귀의 문제

1. 0과 1의 범위를 넘어서는 값이 나올 수 있다.
2. 선형회귀의 추정이 의미가 없을 수 있다. (출력이 확률값으로 확정되어있지 않다)
   우선 타겟값이 두 개의 카테고리로 나누어지는 이진 분류 문제를 생각해보자.
   로지스틱 회귀에서는 어떤 default category에 대한 확률로 Y를 표현한다.

   $$\Pr(\texttt{default} = \texttt{Yes}|\texttt{balance}).$$

   확률값이니, Y는 0에서 1사이의 값을 가진다. 이러한 이진분류 로지스틱 회귀에서는, 어떤 입력이 들어오든 default category에 대한 확률을 예측한다.
   1에서 Y값을 뺀 것이 다른 category에 대한 확률인 것이다.
   아래 그래프를 보면, sigmoid 함수를 이용한 오른쪽 그래프로는 확률값으로서 출력을 사용할 수 있다.

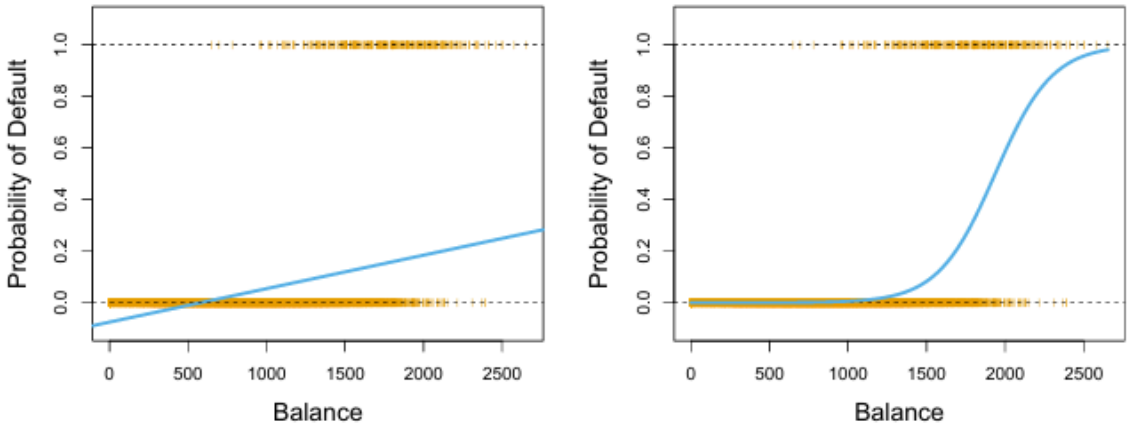

FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default` *(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

for any individual for whom $p(\texttt{balance}) > 0.5$. Alternatively, if a company wishes to be conservative in predicting individuals who are at risk for default, then they may choose to use a lower threshold, such as $p(\texttt{balance}) > 0.1$.

4.3.1 Logistic Model

확률을 Y값으로 설정해 분류 문제를 수행하는 건 좋지만, X와 Pr(X) 사이의 관계는 어떻게 규정할 수 있을까? 전 장에서 배운 linear regression을 그대로 이용할 수 있다.

$$p(X) = \beta_0 + \beta_1 X.$$

하지만 이 선형함수는 y가 0부터 1사이의 확률값을 가져야한다는 constraint에 만족하지 못한다. 이 문제를 해결하기 위해, 우리는 p(X)에 하나의 함수를 더 사용해 만족하게 만든다. 이진 분류에서는 sigmoid function이다.

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

이 함수를 살펴보면, e의 지수에 있는 선형식의 크기가 얼마나 커지던지 작아지던지 항상 0과 1 사이의 출력을 유지한다.

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}.$$

위 식을 조금만 변형시켜보면, 다음과 같은 수식을 얻을 수 있는데, 이 값은 odds라고 부르며, 어떤 일이 일어날 확률을 일어나지 않을 확률로 나누는 것이다. odds는 0부터 무한대까지의 값을 가질 수 있다.
우항 지수함수의 값이 커질 수록 default에 대한 확률값이 더 커질 것이다. odds는 전통적으로 확률값 대신으로 사용하기도 했다.
위 식에 자연로그를 씌우면 이해가 더 쉬울 것이다.

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

즉, 처음 만들었던 선형식이 커질수록 default에 대한 확률값이 커진다는 관계를 명확하게 보일 수 있다.
하지만 여기서 유의해야 할 점은, p(X)는 X에 대한 직선이 아니라는 점이다. p(X)는 비선형적인 sigmoid function이 추가되어 unit한 변화를 보이지 않는다.

4.3.2 - Estimating the Regression Coefficient

공식에 변형을 가했으니, 계수를 추정하는 방식도 변화가 생겼다.
회귀 문제에서는 계수를 추정하기 위해 least squares 방식을 사용했지만, 로지스틱 회귀에서는 이 방식을 사용하지 않는다. 정확히 말하면 사용할 수는 있지만 이 방법은 분류 문제에 적합하지 않다. 출력이 비선형적이고 0부터 1사이의 확률값을 가지기 때문에, least squares 방식을 사용하면 정확한 추정이 불가능하다.
그래서 사용하는 수학적 방법론이 maximum likelihood이다. 이 방법론은 default인 X의 출력이 최대한 1에 가깝도록, 그렇지 않은 X들이 최대한 0에 가깝도록 계수를 추정하도록 한다.

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})).$$

위 함수를 likelihood function이라고 하는데, 계수는 위 함수를 최대화하도록 조정된다.

4.3.4 - Multiple Logistic Regression

이제까지는 단변량 문제를 다뤘다면, 이제 다변량 문제로 문제를 확장해보겠다. 변수가 여러 개일 때 확장하는 방법은 다변량 회귀에서와 동일하다.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−10.8690	0.4923	−22.08	<0.0001
`balance`	0.0057	0.0002	24.74	<0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	−0.6468	0.2362	−2.74	0.0062

- 이상한 점: student로 단변량 로지스틱 회귀를 수행했을 때는 분명 계수가 양수였는데, 왜 음수로 바뀌었을까?
  다른 변수들간의 관계 때문이다.

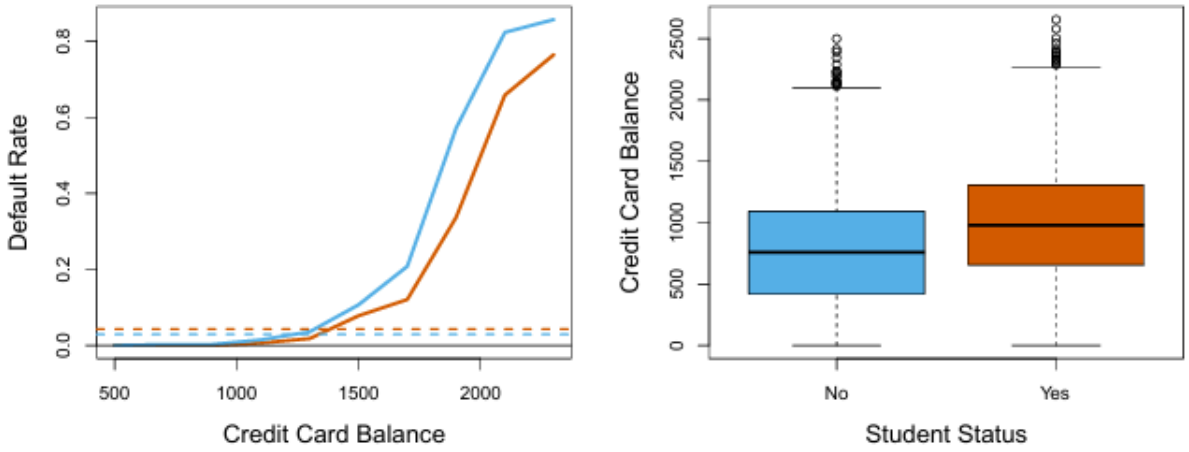

FIGURE 4.3. *Confounding in the* `Default` *data.* Left: *Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of* `balance`*, while the horizontal broken lines display the overall default rates.* Right: *Boxplots of* `balance` *for students (orange) and non-students (blue) are shown.*

이 그래프를 분석해보면, 같은 credit balance 값에 대해서 학생일 때가 default rate가 더 낮

다. 그리고 student가 non-student보다 평균적으로 가지고 있는 credit card balance가 더 높다는 것을 알 수 있다. credit card balance는 양의 계수를 가지고 있었음으로, student 단변량 로지스틱 회귀에서 양의 계수가 나왔다는 것을 알 수 있다. 하지만, 이 두 가지를 변수로 동시에 사용할 때, 오히려 같은 값의 credit balance일 때 학생일 때가 default rate가 적었음으로 음의 coefficient를 가지게 되는 것이다.
이렇게 여러 변수를 동시에 사용하면 기존에 보지 못한 insight를 얻을 수 있다.

4.3.5 Multinomial Logistic Regression

- 이제 class가 2개 이상인 경우에 대해서 문제를 어떻게 설정할 수 있을지 살펴보자.
- 우선 baseline class를 하나 설정한다.

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

or $k = 1, \ldots, K-1$, and

$$\Pr(Y = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}.$$

t is not hard to show that for $k = 1, \ldots, K-1$,

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = K|X = x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p.$$

- baseline을 바꿔서 설정할 수 있기 때문에, 추정과 해석에 있어서 변동성이 클 수 있다.
- baseline이 아닌 classes의 가중치가 baseline과의 관계로 표현된다.
- 즉 baseline 선택에 따라서 달라질 수 있다.
- odds 값도 baseline에 대해 표현하면, class에 대해 배타적인 관계밖에 표현하지 못한다.
- - 이 문제를 해결하기 위해, softmax를 사용한다.
- softmax는 baseline을 설정하지 않고 모든 class에 대한 선형식을 가지고 있다.
- 이제 log odds를 더 나은 방식으로 표현할 수 있다.