ISL
4.3.4~4.3.5

AI명예학회

SKHU

4.3.4- Multiple Logistic Regression

다변량 문제로 쉽게 문제를 확장할 수 있다.
선형 회귀 문제와 동일하게, 변수마다 coefficient를 부여하면 된다.

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

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

AI명예학회
SKHU

4.3.4- Multiple Logistic Regression

- 분명 student(binary)에 대한 coefficient는 단변량 로지스틱 회귀를 수행했을 때 양수였는데, 왜 음수로 바뀌었을까?

-> 다른 변수들과의 관계를 반영했기 때문이다.

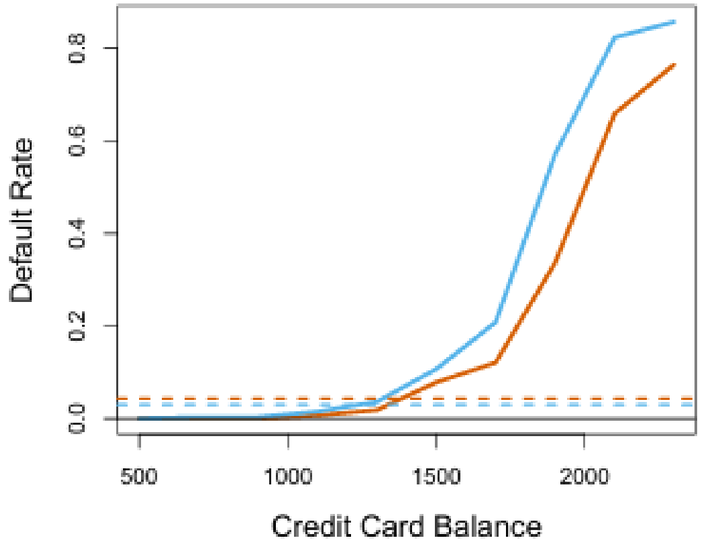
	Coefficient	Std. error	z-statistic	p-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

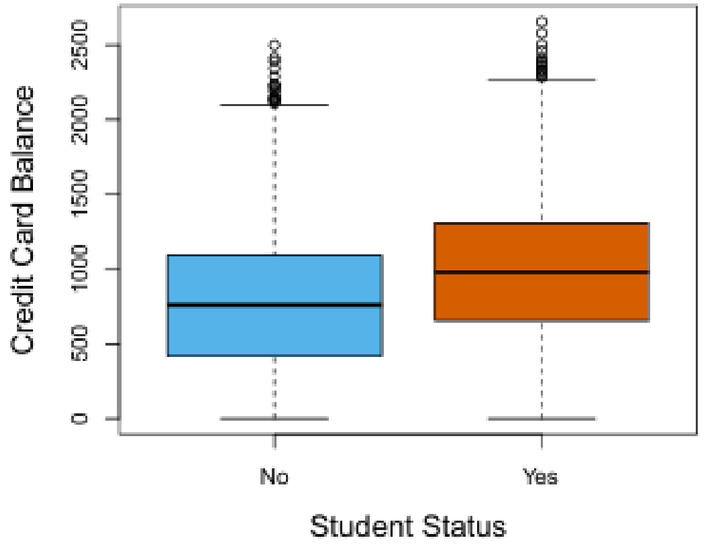
	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

AI명예학회
SKHU

4.3.4- Multiple Logistic Regression

- student가 non-student보다 평균적으로 가지고 있는 credit card balance가 더 높다는 것을 알 수 있다.
- 평균적으로 credit card balance가 더 커서 단변량에서 양의 계수가 나왔다는 것을 알 수 있다.
- 같은 값의 credit balance일 때 학생일 때가 default rate가 적었음으로 음의 coefficient를 가지게 되는 것이다.





AI명예학회
SKHU

4.3.4- Multinomial Logistic Regression

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

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

$: 1, \ldots, K-1$, and

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

$\rm t$ hard to show that for $k = 1, \ldots, K-1$,

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

AI명예학회
SKHU

4.3.4- Multinomial Logistic Regression

- 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.

$$\frac{\Pr(Y = \texttt{stroke} | X = x)}{\Pr(Y = \texttt{epileptic seizure} | X = x)}$$

AI명예학회
SKHU

4.3.4- Multinomial Logistic Regression

- 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.

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

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)}\right) = (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + \cdots + (\beta_{kp} - \beta_{k'p})x_p. \quad (4.14)$$

AI명예학회
SKHU

odds ratio란?

- 실험군의 odds를 대조군의 odds로 나눈 것
- ex) 흡연과 폐암의 관계에 대한 연구일 때, 폐암 환자의 odds를 비환자의 odds로 나눈다.
- odds ratio가 1이면 요인과 관계 없음.
- odds ratio가 1보다 크면 요인에 의해 발생 확률이 큼.
- odds ratio가 1보다 작으면 발생 확률이 작음.

환자-대조군 연구	환자군	대조군
과거 노출	a	b
과거 비노출	c	d

$$\frac{\text{환자군이 노출되었을 오즈}}{\text{대조군이 노출되었을 오즈}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

AI명예학회
SKHU