Classification
(Logistic Regression)

Classification (복습)

- **Response 변수 Y가 카테고리(클래스) 타입일 때를 Classification이라고, 예)**
  - **Y =** 이메일이 스팸인지 아닌지 ∈ { 스팸, 햄 }
  - **Y =** 카드 연체 여부 ∈ { 연체, 상환 }

- **Often we are more interested in estimating the probabilities that X belongs to each category in Response**

Example : 신용카드 Default (채무 불이행) 예측

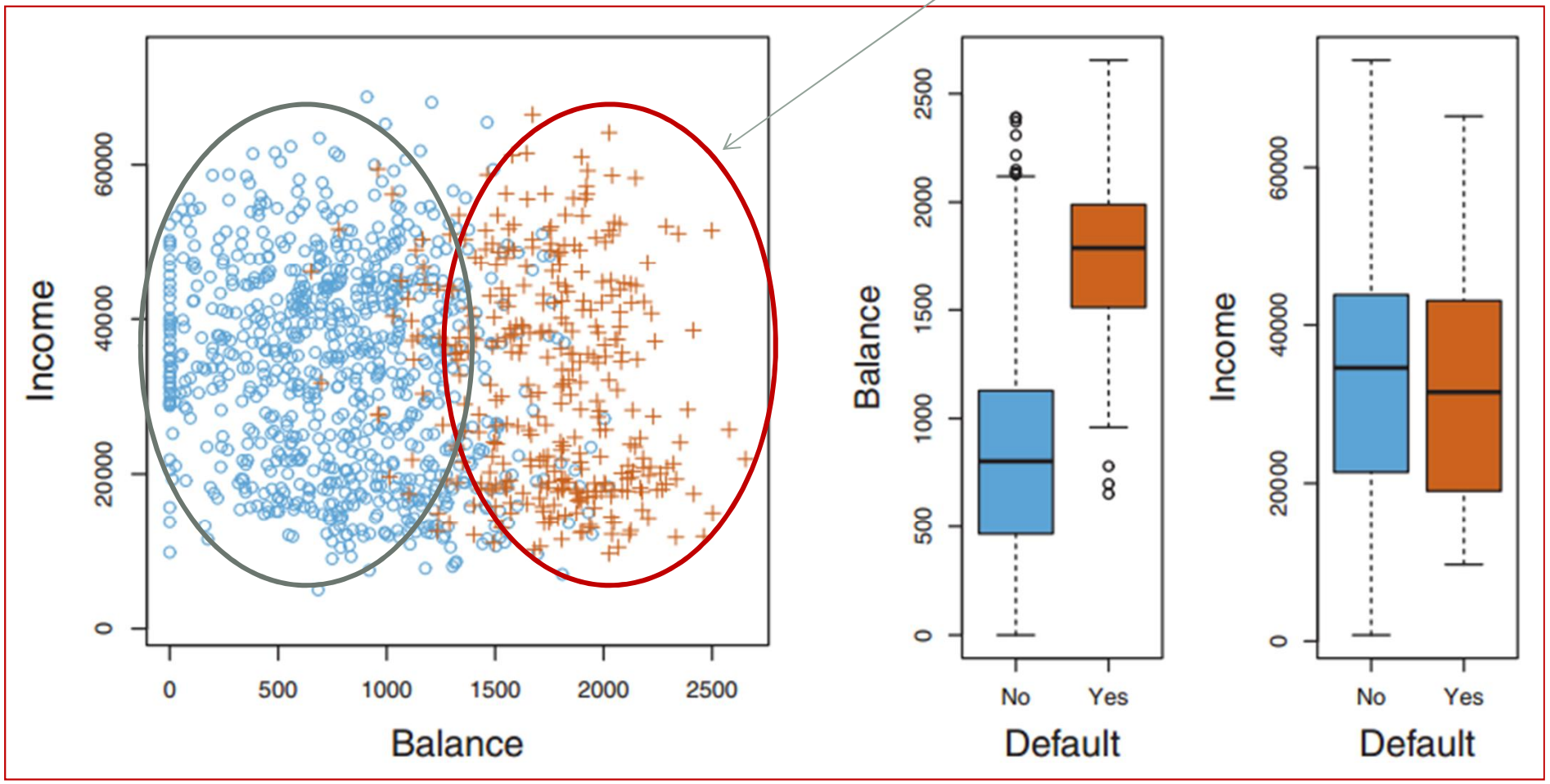
- “어떤 고객이 돈을 갚지 않을까”를 예측하고 싶다

- X variables :
  - Annual Income
  - Monthly credit card balance
  - 학생인가 아닌가

- Response Y 는 Default(연체)를 나타내는 categorical 변수로, No(연체 안함) 또는 Yes(연체) 값 중 하나를 갖는다

- How do we predict the relationship between X and Y?

Example : 신용카드 연체…



- **Response** 변수 **Y**가 카테고리(클래스) 타입인 연체 여부. **Y**가 갖을 수 있는 값은 **Yes** 또는 **No**이다. 입력 **predictor**는 **income**(수입)과 예금잔고(**balance**).
- 위의 **scatter plot**과 **box plot**을 보니, **default**를 예측함에 **Income**은 전혀 도움이 안됨을 금방 알 수 있다. 따라서 **Balance**를 사용하도록 하자.

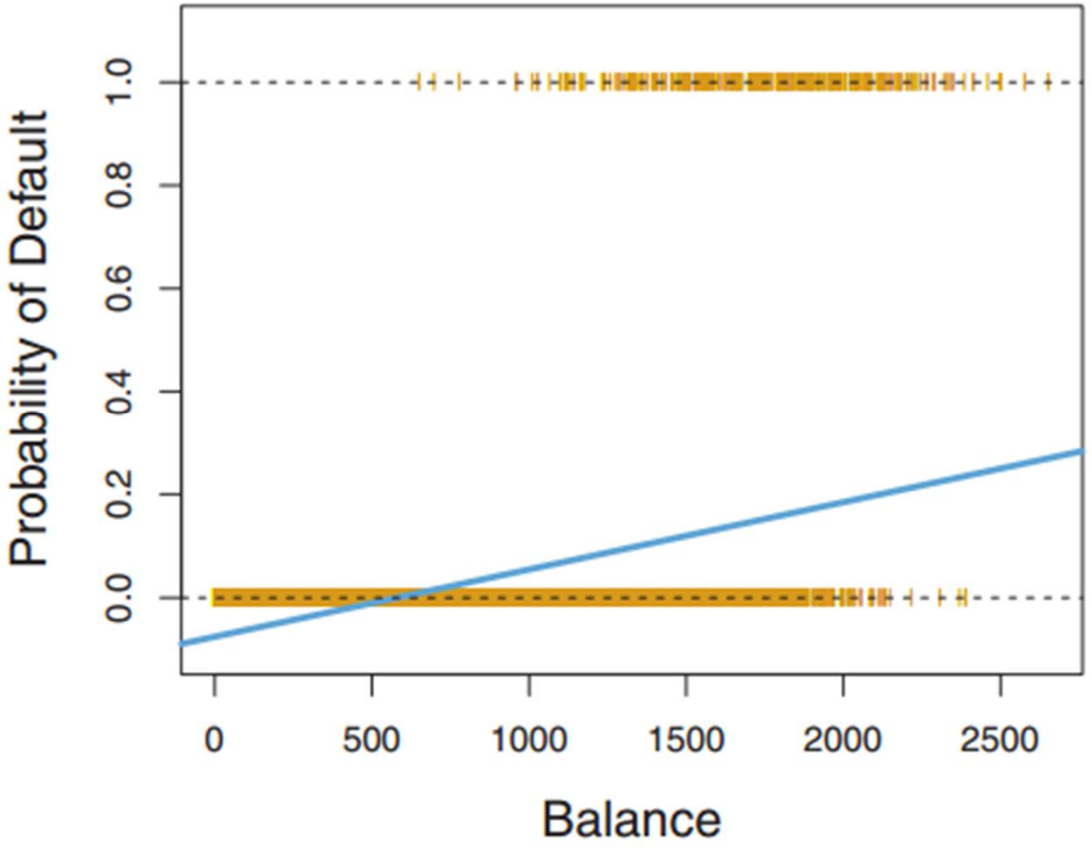
Classification에 Linear Regression을 쓸 수는 없을까?

- 문제 있음

- 카테고리형 **response** 변수의 값이 **{**빨강, 노랑, 파랑**} 3** 개라면 이를 숫자로 어떻게 표현하면 좋을까? 순서를 매기면 되나? 어떤 문제가 있나?
- 카테고리형 **response** 변수의 값이 **Binary 2**개라면? 가령 상환을 **0**, 연체를 **1**로 나타내면? 그래서 추정한 **Y** 값이 **0.5**보다 크면 연체로 분류하고. 그럴 듯 한데... 그런데...

→ 이 경우에도 문제가,

- Balance가 500 이하가 되니 추정한 Y값이 0 보다 작음.
- Balance가 아주 크면 1보다 큰 값이 나올텐데 어떻게 해야 하나?
- Balance의 범위가 0에서 3000을 넘지 않는 것 같이 보이는데, 그럼 “probability of default”가 0.5 아래에 다 있겠네.



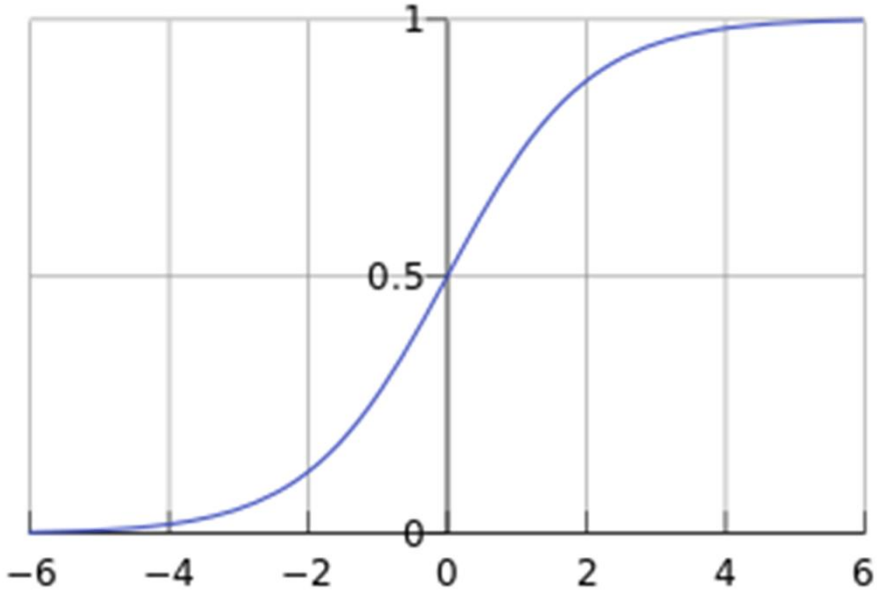
솔루션 : Logistic Regression (이름은 'regression'이나 실제는 Classification)

- Linear regression으로 [0, 1] 사이의 값을 target response로 직접 추정하려고 하지 말고, 고객이 연체(Y=1) 할 확률을 구하는 방법을 택하자.

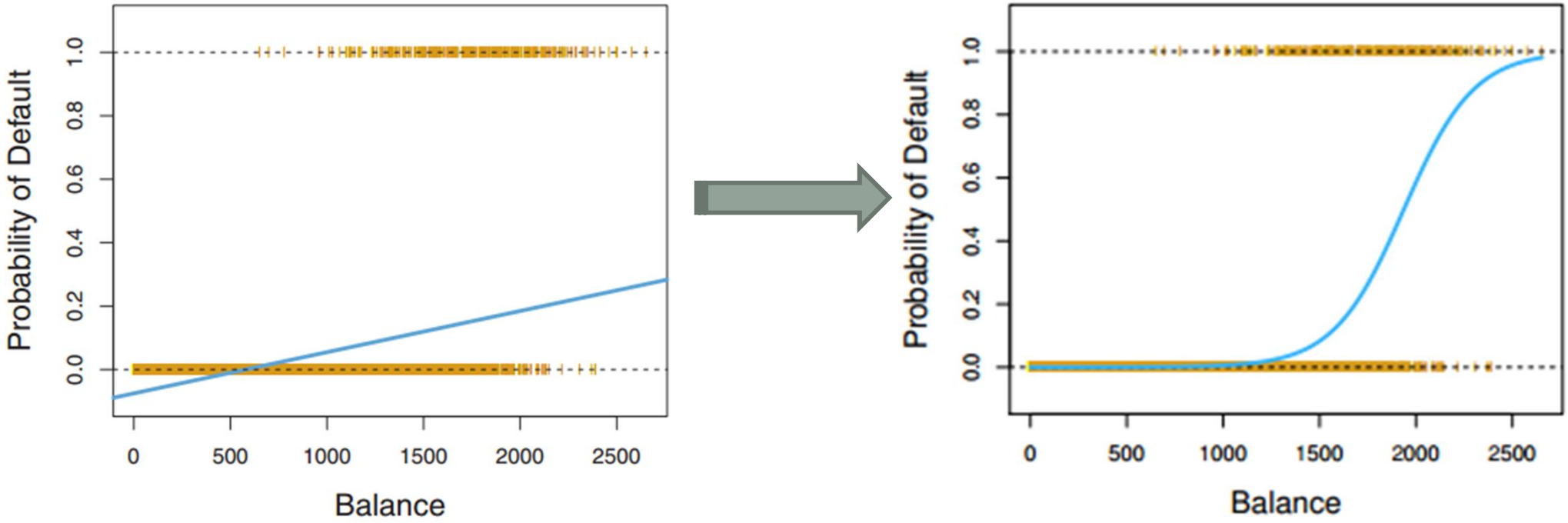
  즉, Prob(response Y = 1 | X)를 구하자
- Let's model P(Y = 1| X) using a function that gives outputs between 0 and 1.
- Use Logistic Regression :

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

$0 < p < 1$



Logistic Regression 을 사용하면,



- Balance(잔고)가 아주 작더라도 연체확률이 0이하로 되지 않고, 아주 크더라도 1보다 더 커지지 않는다. 잔고가 2000 정도면 연체확률이 0.5 정도가 된다. p 값에 따라 고객을 연체 또는 상환 중 어디에 분류할까는 상황에 맞게 정하면 된다.

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

➢ 위의 ***p*** 에 **'default'**한 **X**들을 넣으면 가능한 **1**에 가까운 수가 나오고, **'default'**하지 않은 **X**들을 넣으면 가능한 **0**에 가까운 수가 나오도록 만드는 그런 β_1, β_0 를 구하자

■ MLE를 사용 Logistic Regression 의 Coefficient 구하기 :

$$\ell(\beta_0, \beta_1) = \prod_{i} p(x_i) \prod_{i'} (1 - p(x_{i'})).$$ **Likelihood Function**

This *likelihood* gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data. **: Maximum Likelihood**

Most statistical packages can fit linear logistic regression models by maximum likelihood. In `R` we use the `glm` function.

β_0 , β_1 :

	Coefficient	Std. Error	Z-statistic	P-value
β_0 : Intercept	-10.6513	0.3612	-29.5	< 0.0001
β_1 : balance	0.0055	0.0002	24.9	< 0.0001

- Interpreting what β_1 means is not very easy with logistic regression, because we are predicting P(Y) and not Y
- If β_1 =0 : Y 와 X 가 관계없다.
- If β_1 >0 : X가 커지면 Prob(Y=1) 또한 커짐
- If β_1 <0 : X가 커지면 반대로 Prob(Y=1) 는 작아짐
- 구체적으로 P(Y) 가 X가 1 unit 증가함에 따라 얼마나 변할 지는 (P(Y)가 X에 대해 linear하지 않으므로) 현재 X의 값에 따라 달라진다. 하지만 현재 X가 얼마이건 β_1 이 양수이기에 P(Y)가 증가함은 맞다.

결과 해석

	Coefficient	Std. Error	Z-statistic	P-value
β_0 : Intercept	-10.6513	0.3612	-29.5	< 0.0001
β_1 : balance	0.0055	0.0002	24.9	< 0.0001

- Z-statistic이 linear regression의 t-statistic 역할을 한다. Z-statistic의 **절대값**이 클수록 null hypothesis $\beta_1 = 0$ 를 거부하기 좋다.
- Z-statistic 이 크고(P-value가 작아) null hypothesis를 거부할 수 있고 따라서 이 데이터에 의하면 response인 'default'와 predictor인 'balance'간에 연관이 있다고 결론지을 수 있다.
- 이 때 β_1 이 positive 이므로 balance가 커지면 연체할 확률 p 가 증가함을 추정할 수 있다.

Making Predictions

- Balance(잔고)가 $1000 인 사람의 연체 확률은?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.00576 < 1\%$$

- Balance(잔고)가 $2000 인 사람의 연체 확률은?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586 = 58.6\%$$

- 확률은 위와 같이 나오지만 이들을 어떻게 분류할까는 사정에 따라 달라질 수 있다. 가령 연체로 인한 피해가 매우 크면 연체 가능자에 대해 몹시 신중하게 볼 것이다. 이 경우 연체 확률이 10%만 되어도 “연체”로 분류할 수도 있다

Qualitative Predictors in Logistic Regression

- Logistic regression의 predictor가 qualitative variable, 즉 카테고리형 변수이면 linear regression에서와 같이 그 변수를 dummy 변수화해 모델을 만들 수 있다.
- Student 변수는 카테고리형이므로 이를 dummy 변수화 하여 (Student = 1, Non-student =0).

	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

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

- β_1 이 양수이기에 학생은 학생이 아닌 경우보다 더 연체할 가능성이 많다.

Multiple Logistic Regression

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

where $X = (X_1, \ldots, X_p)$ are p predictors

- Predict Default using :
  - ✓Balance (quantitative)
  - ✓Income (quantitative)
  - ✓Student (qualitative/카테고리형)

	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

Predictions

- A student with a credit card balance of $1,500 and an income of $40,000 has an estimated probability of default of

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}} = 0.058$$

- A non-student with a credit card balance of $1,500 and an income of $40,000 has an estimated probability of default of

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}} = 0.105$$

- 학생-비학생만 놓고 보았을 때는 학생이 더 연체할 확률이 높았는데, 다른 변수들과 함께 놓고 보니 오히려 다른 조건이 같다면 학생이 연체를 적게한다.

Confounding

	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

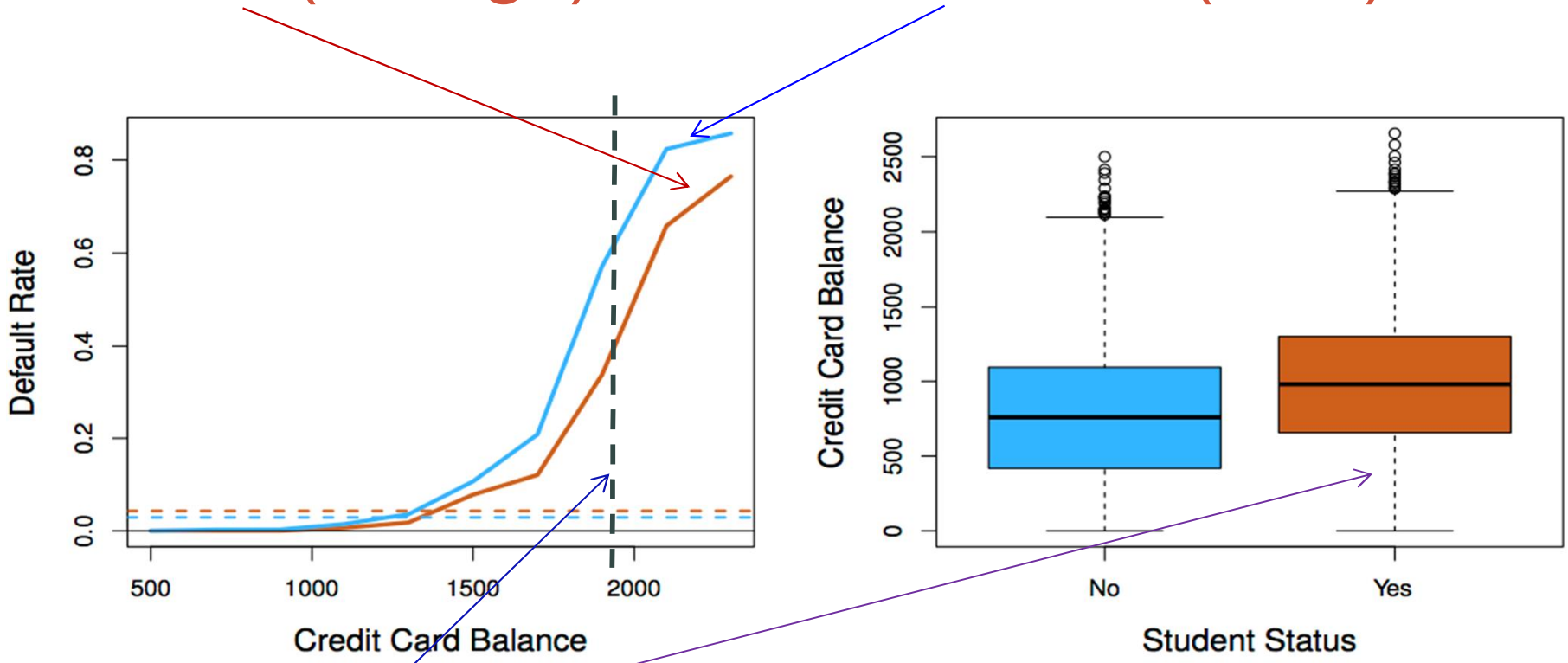
Positive

	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

Negative

- Multiple logistic regression에서 balance와 student의 p-value가 이들이 p(X)와 관련있음을 가르킨다. 그런데, 앞에서는 student[Yes]의 coefficient가 양수이어서 student이면 연체의 확률이 증가했는데, multiple logistic regression에서는 student[Yes] coefficient 가 음수가 되었다. 어찌된 것일까?

Students (Orange) vs. Non-students (Blue)



- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students.
- But for each level of balance, students default less than non-students.
- Multiple logistic regression can tease this out

Logistic Regression for >2 Response Classes

- Response 카테고리가 { 파란눈, 갈색눈, 까만눈} 같이 3개 이상이어도 logistic regression 적용이 가능하다

- 이 경우, Pr(Y=파란눈 | X)와 Pr(Y=갈색눈 | X)를 구하면 나머지 Pr(Y=까만눈 | X) 은 별도로 구할 필요없이 1 - Pr(Y=파란눈 | X) - Pr(Y=갈색눈 | X) 로 구할 수 있다