Logistic Regression: Interpretation

강필성
고려대학교 산업경영공학부
pilsung_kang@korea.ac.kr

AGENDA

Logistic Regression: Interpretation

- Meaning of coefficients
  - ✓ Linear regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$

    - The amount of target variable changes when the input variable is increased by 1
  - ✓ Logistic regression

$$log(Odds) = log\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$

$$p = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d)}}$$

    - The amount of log odd changes when the input variable is increased by 1 (not intuitive)

Logistic Regression: Interpretation

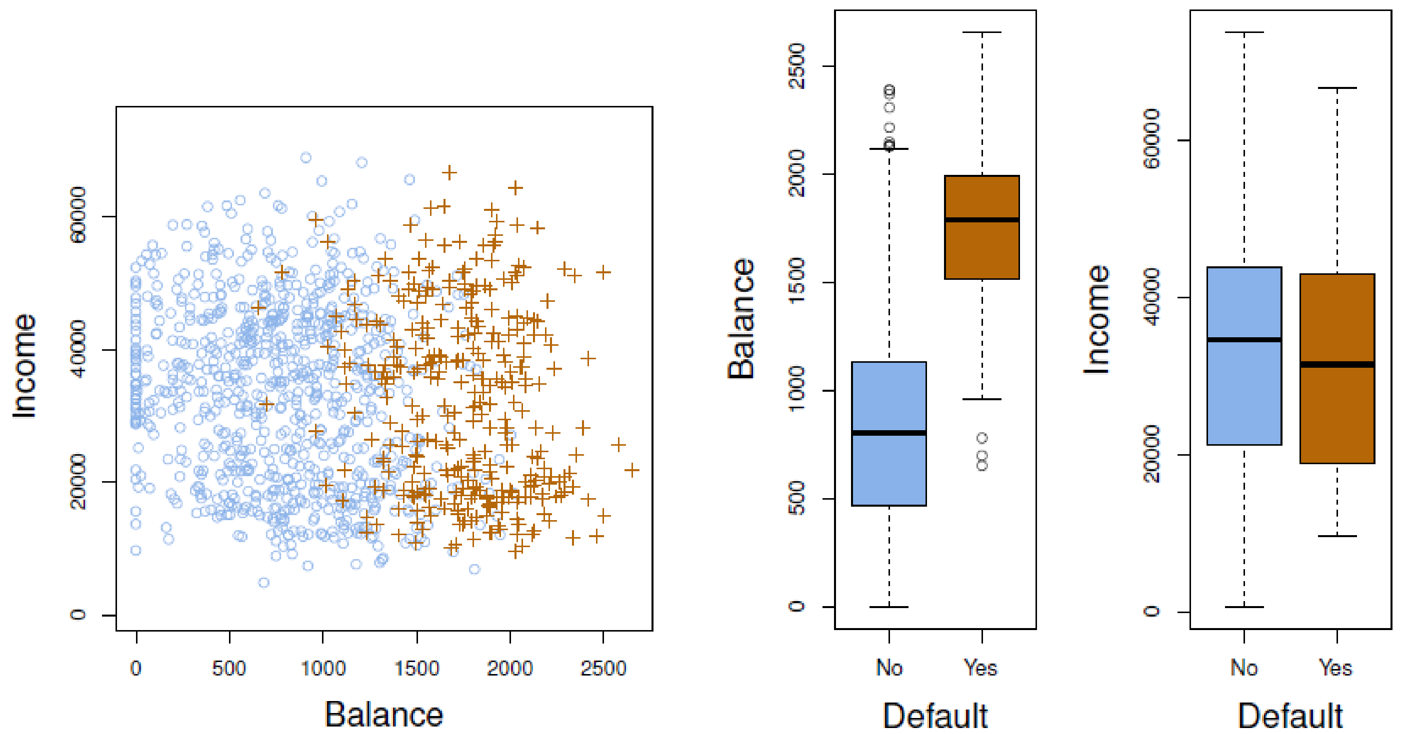
- Odds ratio
  - ✓ Suppose that the value of x_1 is increased by one unit from x_1 to x_1+1, while the other predictors are held at their current value.
  - ✓ Odds ratio:

$$\frac{odds(x_1+1,\cdots,x_d)}{odds(x_1,\cdots,x_d)} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1(x_1+1)+\hat{\beta}_2x_2+\cdots+\hat{\beta}_dx_d}}{e^{\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_dx_d}} = e^{\hat{\beta}_1}$$

  - ✓ When x_1 is increased by 1, then the odds is increased(decreased) by a factor of $e^{\hat{\beta}_1}$
    - ▪ Coefficient is positive → success probability increases when the corresponding input value increases (success class and coefficient are positively correlated)
    - ▪ Coefficient is positive → success probability increases when the corresponding input value increases (success class and coefficient are negatively correlated)

Logistic Regression: Example 1

- Credit Card Default



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

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

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

Logistic Regression: Example 1

- Credit Card Default: single variable

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

What is our estimated probability of default for someone with a balance of \$1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

With a balance of \$2000?

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

Logistic Regression: Example 1

- Credit Card Default: multiple variables

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

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

	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

Logistic Regression: Example 2

- Personal Loan Offer
  - ✓ Predict a new customer whether he/she will accept the bank's personal loan offer

일련 번호	나이	경력	소득	가족 수	월별 신용카드 평균사용액	교육 수준	담보부 채권	개인 대출	증권 계좌	CD 계좌	온라인 뱅킹	신용 카드
1	25	1	49	4	1.60	UG	0	No	Yes	No	No	No
2	45	19	34	3	1.50	UG	0	No	Yes	No	No	No
3	39	15	11	1	1.00	UG	0	No	No	No	No	No
4	35	9	100	1	2.70	Grad	0	No	No	No	No	No
5	35	8	45	4	1.00	Grad	0	No	No	No	No	Yes
6	37	13	29	4	0.40	Grad	155	No	No	No	Yes	No
7	53	27	72	2	1.50	Grad	0	No	No	No	Yes	No
8	50	24	22	1	0.30	Prof	0	No	No	No	No	Yes
9	35	10	81	3	0.60	Grad	104	No	No	No	Yes	No
10	34	9	180	1	8.90	Prof	0	Yes	No	No	No	No
11	65	39	105	4	2.40	Prof	0	No	No	No	No	No
12	29	5	45	3	0.10	Grad	0	No	No	No	Yes	No
13	48	23	114	2	3.80	Prof	0	No	Yes	No	No	No
14	59	32	40	4	2.50	Grad	0	No	No	No	Yes	No
15	67	41	112	1	2.00	UG	0	No	Yes	No	No	No
16	60	30	22	1	1.50	Prof	0	No	No	No	Yes	Yes
17	38	14	130	4	4.70	Prof	134	Yes	No	No	No	No
18	42	18	81	4	2.40	UG	0	No	No	No	No	No
19	46	21	193	2	8.10	Prof	0	Yes	No	No	No	No
20	55	28	21	1	0.50	Grad	0	No	Yes	No	No	Yes

Logistic Regression: Example 2

- Data Preprocessing

  - A total of 5,000 customers
  - Predictors
    - ✓ Demographic: age, income, etc.
    - ✓ Relationship with the bank: mortgage, security account, etc.
  - Only 480(9.6%) accepted the personal loan.

  - 60% for training, 40% for validation.
  - Create dummy variables for the categorical predictors.

$$EducProf = \begin{cases} 1 \text{ if education is } \textit{Professional} \\ 0 \text{ otherwise} \end{cases}$$

$$EducGrad = \begin{cases} 1 \text{ if education is at } \textit{Graduate} \text{ level} \\ 0 \text{ otherwise} \end{cases}$$

Logistic Regression: Example 2

- Modeling with all input variables

$$p = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d)}}$$

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

Logistic Regression: Interpretation

- Coefficient
  - ✓ The beta values for corresponding input variables
  - ✓ The value is the changing ratio of log odds when the input variable increases by 1
  - ✓ Positive value: positively correlated with the success class
  - ✓ Negative value: negatively correlated with the success class

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

Logistic Regression: Interpretation

- p-value
  - ✓ Indicating whether the corresponding input variable is statistically significant or not
  - ✓ Significance is strongly supported when the p-value is close to 0

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

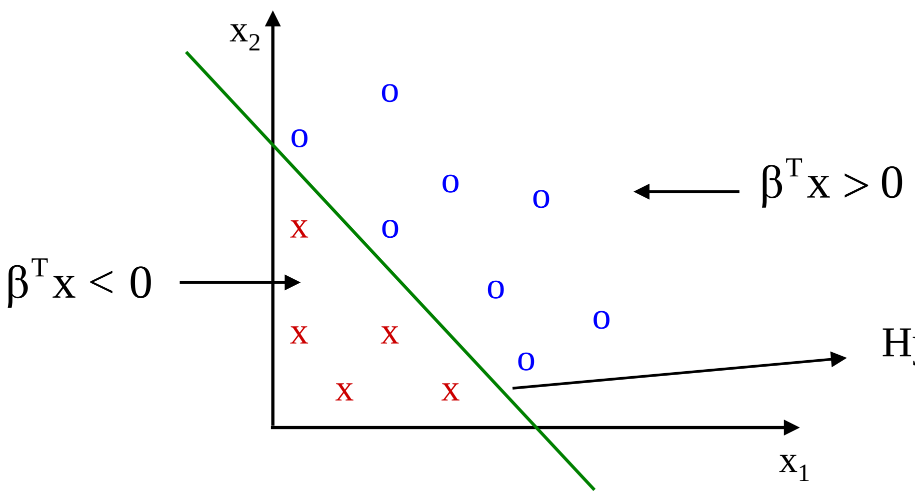
Logistic Regression: Interpretation

- Odds ratio
  - ✓ The ratio of odds when the value of the corresponding input variable increases by 1

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-13.20165825	2.46772742	0.00000009	*
Age	-0.04453737	0.09096102	0.62439483	0.95643985
Experience	0.05657264	0.09005365	0.5298661	1.05820346
Income	0.0657607	0.00422134	0	1.06797111
Family	0.57155931	0.10119002	0.00000002	1.77102649
CCAvg	0.18724874	0.06153848	0.00234395	1.20592725
Mortgage	0.00175308	0.00080375	0.02917421	1.00175464
Securities Account	-0.85484785	0.41863668	0.04115349	0.42534789
CD Account	3.46900773	0.44893095	0	32.10486984
Online	-0.84355801	0.22832377	0.00022026	0.43017724
CreditCard	-0.96406376	0.28254223	0.00064463	0.38134006
EducGrad	4.58909273	0.38708162	0	98.40509796
EducProf	4.52272701	0.38425466	0	92.08635712

Logistic Regression: Interpretation

- Geometric interpretation
  - ✓ Can be thought of as finding a hyper-plane to separate positive and negative data points.



Hyperplane: $\beta^T x = 0$ where $\beta = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$

Classifier

$$y = \frac{1}{\left(1 + \exp(-\beta^T x)\right)}$$

$$\begin{pmatrix} y \to 1 & if & \beta^T x \to & \infty \\ y = \frac{1}{2} & if & \beta^T x = & 0 \\ y \to 0 & if & \beta^T x \to & -\infty \end{pmatrix}$$

Logistic Regression: Interpretation

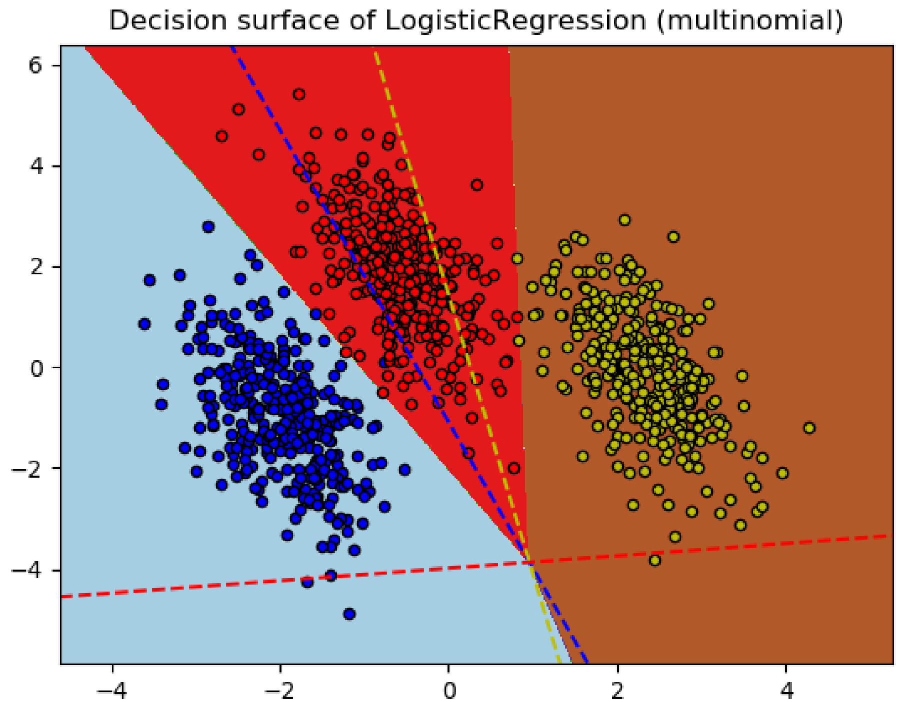
- Profiling
  - ✓ Finding factors that differentiate between the two classes.
  - ✓ After variable selection:

$$\frac{p}{1-p} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d}$$

  - ✓ Variables associated with positive β_i increase the probability of the success.
  - ✓ Variables associated with negative β_i decrease the probability of the success.

Multinomial Logistic Regression

- Basic Logistic Regression is developed to solve the binary classification problem
  - ✓ Q) Can we use the logistic regression to classify more than 3 classes?



http://scikit-learn.org/stable/auto_examples/linear_model/plot_logistic_multinomial.html

Multinomial Logistic Regression

- Multinomial logistic regression
  - ✓ Set the baseline class and formulate the regression equation for the relative log odds to this class
  - ✓ Ex) If there are three classes, estimate the coefficients of the following two regression models
    - Logistic regression of Class 1 versus Class 3

$$log\left(\frac{p(y=1)}{p(y=3)}\right) = \hat{\beta_{10}} + \hat{\beta_{11}}x_1 + \hat{\beta_{12}}x_2 \cdots + \hat{\beta_{1d}}x_d = \boldsymbol{\beta}_{1.}^T\mathbf{x}$$

    - Logistic regression of Class 2 versus Class 2

$$log\left(\frac{p(y=2)}{p(y=3)}\right) = \hat{\beta_{20}} + \hat{\beta_{21}}x_1 + \hat{\beta_{22}}x_2 \cdots + \hat{\beta_{2d}}x_d = \boldsymbol{\beta}_{2.}^T\mathbf{x}$$

Multinomial Logistic Regression

- Multinomial logistic regression
  - ✓ Why do we learn only two models although there are three classes? (Generally, why do we learn (K-1) models when there are K classes?)
    - For each object, the sum of likelihoods must be 1, so that if we know (K-1) likelihoods, that the rest can be automatically computed

$$\frac{p(y=1)}{p(y=3)} = e^{\boldsymbol{\beta}_{1\cdot}^T \mathbf{x}} \qquad \frac{p(y=2)}{p(y=3)} = e^{\boldsymbol{\beta}_{2\cdot}^T \mathbf{x}}$$

$$p(y=1) + p(y=2) + p(y=3) = 1$$

$$p(y=3) \times e^{\boldsymbol{\beta}_{1\cdot}^T \mathbf{x}} + p(y=3) \times e^{\boldsymbol{\beta}_{2\cdot}^T \mathbf{x}} + p(y=3) = 1$$

$$p(y=3) = \frac{1}{1 + e^{\boldsymbol{\beta}_{1\cdot}^T \mathbf{x}} + e^{\boldsymbol{\beta}_{2\cdot}^T \mathbf{x}}}$$

Multinomial Logistic Regression

- Interpreting the coefficients in multinomial logistic regression
  - ✓ Interpret the coefficients for the two compared classes
    - Total phenols, Flavanoids, Monflavanoid penols, Hue, OD280~ variables are statistically significant for both 1 vs. 3, 2 vs. 3 models
    - Ash., Proanthocyanins variable is not statistically significant when discriminating the classes 1 and 3, but is significant when discriminating the classes 2 and 3

		1 vs 3		2 vs 3
	Coefficient	p-value	Coefficient	p-value
(Intercept)	-223.7894	0.0000	340.9326	0.0000
Alcohol.2	19.6193	0.7880	-35.2596	0.6828
Malic.acid.	1.0581	0.9228	-0.3022	0.9899
Ash.	14.6800	0.3881	-204.7437	0.0000
Alcalinity.of.ash.	-20.3881	0.8815	-2.2832	0.9864
Magnesium.	2.0553	0.9975	2.1132	0.9974
Total.phenols.	-169.4205	0.0000	-40.3325	0.0000
Flavanoids.	193.7935	0.0000	16.2013	0.0188
Nonflavanoid.phenols	93.5409	0.0000	214.1837	0.0000
Proanthocyanins.	15.5178	0.1453	115.3184	0.0000
Color.intensity.	-16.6775	0.4212	-11.5066	0.7671
Hue	-50.0008	0.0000	352.7617	0.0000
OD280.OD315.of.diluted.wines.	75.2435	0.0000	84.2914	0.0000
Proline.	-0.0120	1.0000	-0.2899	0.9999


ANY
questions?