Restor Dent Endod. 2017 Nov;42(4):342-348
https://doi.org/10.5395/rde.2017.42.4.342
pISSN 2234-7658·eISSN 2234-7666

Restorative Dentistry & Endodontics RDE

Open Lecture on Statistics

Check for updates

Statistical notes for clinical researchers: logistic regression

Hae-Young Kim *

Department of Health Policy and Management, College of Health Science, and Department of Public Health Science, Graduate School, Korea University, Seoul, Korea

OPEN ACCESS

***Correspondence to**
Hae-Young Kim, DDS, PhD
Professor, Department of Health Policy and Management, College of Health Science, and Department of Public Health Science, Graduate School, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea.
Tel: +82-2-3290-5667
Fax: +82-2-940-2879
E-mail: kimhaey@korea.ac.kr

ORCID iDs
Hae-Young Kim
https://orcid.org/0000-0003-2043-2575

Logistic regression is a regression model where the dependent variable is categorical and corresponding independent variables can be categorical or continuous. This article covers the case of a binary dependent variable such as an event occurring coded 1 = 'event' and 0 = 'no event'. Frequent outcomes are pass/fail, win/lose, disease/no disease, etc. The logistic regression model estimates the probability that an event occurs versus the probability that the event does not occur.

An example: score and pass data

Let's say that an institution performed an assessment procedure to determine pass and fail of the participants considering exam scores, interview result, and reputation among colleagues. **Table 1** shows a data with 2 variables, exam scores and pass state (1 = pass, 0 = fail). We can notice that there is a trend that persons with lower scores are more likely to fail, while persons with higher scores tend to pass. When we plot the data as **Figure 1A**, we can see persons with value 1 (pass) have scores that shift to the right side, while persons with value 0 (fail) have those that shift to the left side. Persons with same score may not have the same outcome (*e.g.*, cases of score = 799) because the assessment procedure comprises other factors. At least we can postulate that the probability of pass may be higher if the score is higher. What is the best-fit line for this data? A usual straight regression line ranging from minus infinity to infinity does not make sense for this case. Instead of ordinal regression the logistic regression can fit the probability more adequately. In **Figure 1B**, the probability estimated by logistic regression is presented. The estimated probability by the logistic regression model (red dot and line) seems reasonable because it reflects the observed reality that the probability of pass decreases close to zero with very low scores, while the probability increases close to one with very high scores.

Table 1. Scores of applicants who passed the final assessments

Score	755	755	763	781	783	788	792	793	798	799	799	802	813	824	845
Pass	0	0	0	0	1	1	0	1	0	0	1	1	1	1	1

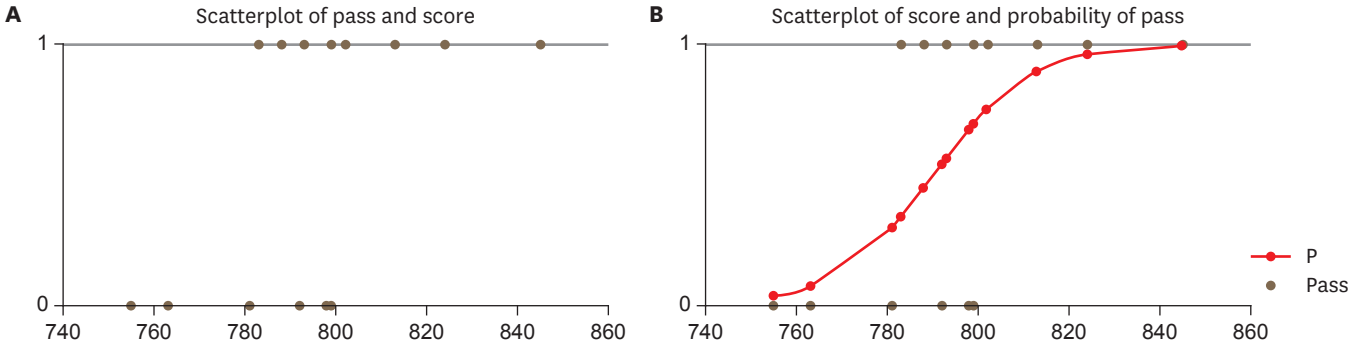


Figure 1. Scatterplot of pass (1 = pass, 0 = fail) and score: (A) pass and score, and (B) estimated probability (P) of pass added.

Review of probability, odds, and odds ratio

From the previous sections about risk, odds, and odds ratio, they were defined as following formulas:

$$\text{Probability or risk (p)} = \frac{\text{number of events}}{\text{number of all observations}}$$

$$\text{Odds} = \frac{\text{p (event)}}{\text{p (non-even)}} = \frac{\text{p}}{1-\text{p}}$$

$$\text{Odds ratio} = \frac{\text{odds}_1}{\text{odds}_0} = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}$$

Let's consider an example of flipping of fair coins vs. loaded coins.

Fair coin flip	Loaded coin flip
$P\text{ (heads)} = \frac{1}{2} = 0.5$	$P\text{ (heads)} = \frac{8}{10} = 0.8$
$\text{Odds (head)} = \frac{0.5}{1-0.5} = \frac{0.5}{0.5} = 1$	$\text{Odds (head)} = \frac{0.8}{1-0.8} = \frac{0.8}{0.2} = 4$
$\text{Odds ratio (loaded } vs. \text{ fair coin flip)} = \frac{\text{odds}_1}{\text{odds}_0} = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \frac{\frac{0.8}{0.2}}{\frac{0.5}{0.5}} = 4$	

Odds ratio is important in interpreting in logistic regression because it represents how much the odds change with 1 unit increase in the predictor variables while keeping all other variables constant.

Logistic regression

1. Logit link function

Logistic regression uses logit link function to estimate unknown probability of outcome (p) for a linear combination of predictor variables. The original probability ranging from zero to one cannot match with linear combination of predictor variables ranging minus infinity to infinity [1].

$$\text{Logit (p)} = \ln\text{ (odds)} = \log_e\left(\frac{\text{p}}{1-\text{p}}\right)$$

where $\log_e x = \ln(x)$ and e = Euler's number, 2.71828.

https://doi.org/10.5395/rde.2017.42.4.342

Table 2. Logit transformation from probability (p)

P	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99
1–p	0.99	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.01
odds	0.01	0.11	0.25	0.43	0.67	1.00	1.50	2.33	4.00	9.00	99.00
Logit (p) = ln (odds)	−4.60	−2.20	−1.39	−0.85	−0.41	0.00	0.41	0.85	1.39	2.20	4.60

Logit link function accommodate p ranging from zero to one. The logit link function reconciles the incongruity by changing the range of dependent variable, p, into minus infinity to infinity. As seen in **Table 2**, final logit (p) values cover from minus values to plus values.

2. Property of logit and inverse logit

Shown in **Figure 2A**, logit function has an s-shaped curve. Logit (p) is undefined at p = 0 and p = 1. When p approaches close to zero, the value of logit (p) goes toward minus infinity and when p get larger close to one, it goes toward infinity. We can notice that the logit (p) has a value of zero at p = 0.5.

Figure 2B shows inverse logit graph. Inverse logit returns the probability of the event ranging from zero to one. **Figure 1B** and **Figure 2B** show similar shape because both represent estimated probability. The induced inverse logit formula is as following:

$$\text{Inverse logit (p)} = \log^{-1}\left(\frac{p}{1-p}\right) = \frac{1}{1+e^{-\alpha}} = \frac{e^{\alpha}}{1+e^{\alpha}} = p$$

where α = some number.

3. Estimation of logistic regression equation

Simple logistic regression is expressed as logit (p) and linear combination of predictor variables as below.

$$\text{logit (p)} = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

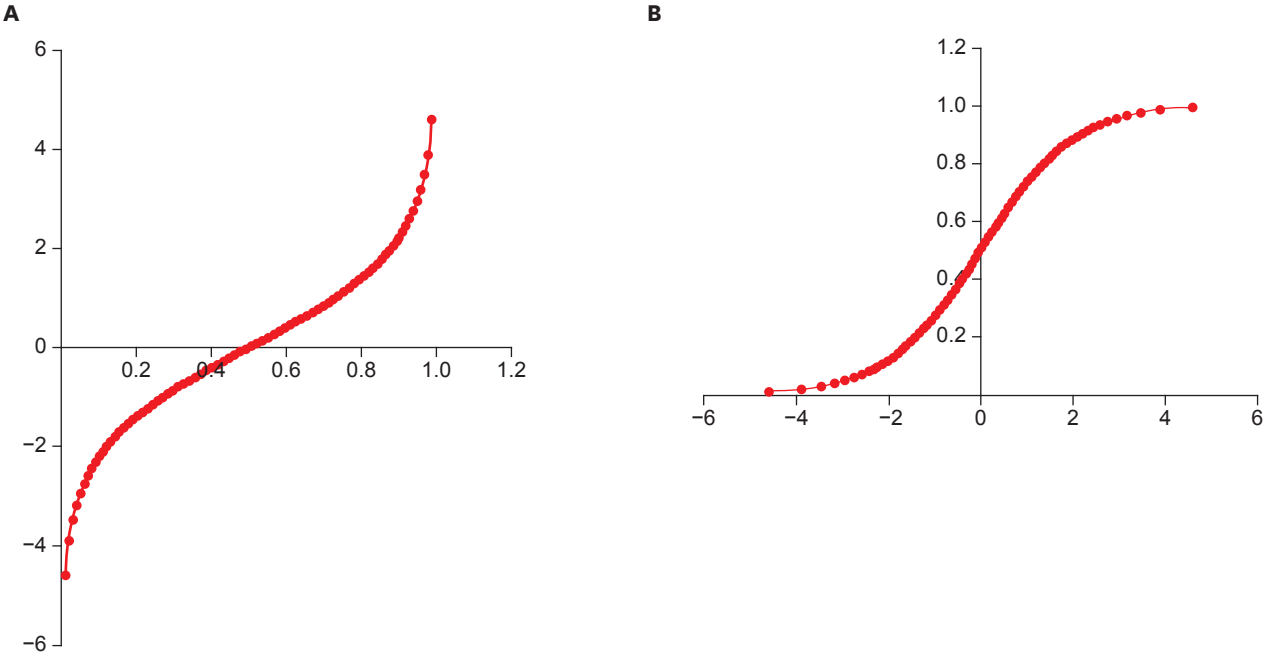


Figure 2. Probability and logit transformation: (A) natural log of adds ratio (logit [p]), (B) inverse logit (p).

https://doi.org/10.5395/rde.2017.42.4.342

Using a fictitious data based on the example above logistic regression was performed and the output was provided (pages 348–349). The observations (n = 15) are multiplied by 100 to provide high power to get significant estimates artificially. The dependent variable was the binary variable pass and score was the predictor variable. The SPSS (IBM Corp., Armonk, NY, USA) output of (e) below gives coefficients as following.

<SPSS output>

β_0 = −73.578140 (p<0.001), β_1 = 0.093115 (p<0.001)

Exp (β_1) = 1.098, 95% confidence interval of Exp (β_1) = [1.086, 1.109]

The estimated logistic equation is:

$$\text{logit}\,(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x = -73.578140 + 0.093115 \times (\text{Score})$$

where p = probability of 'pass'.

Here represents odds ratio which means the amount of change in odds with 1 unit increase in the predictor variable. The odds ratio, exp (β_1) = $e^{0.093115}$ = 1.097588. Therefore, as the score increases by 1 point, the odds of pass was estimated to increase by 9.8%. The 95% confidence interval of odds ratio was [1.086, 1.109] which does not include a value one. Odds ratio value of one means that 1 unit increase in the predictor variable does not make any difference in odds. Therefore, to get statistical significance, it is important to confirm that 95% confidence interval of odds ratio does not include one.

1) Estimated probability

After some algebra, inverse logit gives us the estimated probability by the predictor variable as follows:

$$\text{logit}\,(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x} \rightarrow p = e^{\beta_0+\beta_1 x}(1-p) \rightarrow p(1+e^{\beta_0+\beta_1 x}) = e^{\beta_0+\beta_1 x} \rightarrow \hat{p} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

To get the probability of pass at score 781, we can use the estimated probability function. Also, if the score increases by one point to 782 then the estimated probability can be calculated as shown in **Table 3**. According to the results for the score 781, estimated

Table 3. Estimated probability and odds ratio based on logistic regression model

Score = 781	$\hat{p} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{e^{-73.578140+0.093115(781)}}{1+e^{-73.578140+0.093115(781)}} = \frac{e^{-0.85533}}{1+e^{-0.85533}} = \frac{0.425145}{1+0.425145} = 0.298317$ $\text{odds} = \frac{p}{1-p} = \frac{0.298317}{1-0.298317} = 0.425145$
Score = 782	$\hat{p} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{e^{-73.578140+0.093115(782)}}{1+e^{-73.578140+0.093115(782)}} = \frac{e^{-0.76221}}{1+e^{-0.76221}} = \frac{0.466634}{1+0.466634} = 0.318167$ $\text{odds} = \frac{p}{1-p} = \frac{0.318167}{1-0.318167} = 0.466634$
Odds ratio for a 1 point increase in score:	Odds ratio for a 1 point increase in score: $\frac{\text{odds at 782}}{\text{odds at 781}} = \frac{0.466634}{0.425145} = 1.097588$

https://doi.org/10.5395/rde.2017.42.4.342

Table 4. Scores, estimated probabilities, and odds ratios based on logistic regression model

Score	755	755	763	781	783	788	792	793	798	799	799	802	813	824	845
P	0.04	0.04	0.07	0.30	0.34	0.45	0.54	0.57	0.67	0.69	0.69	0.75	0.89	0.96	0.99
Odds	0.04	0.04	0.08	0.43	0.51	0.82	1.18	1.30	2.07	2.27	2.27	3.01	8.37	23.31	164.84

probability of pass in the assessment is 0.30 or 30%. Also, the odds ratio is obtained as 1.098, which is the same value with exp (β_1) from the SPSS output, representing the increase of odds of 9.8% related to a 1 point increase of the score.

Estimated probability for other score values are shown in the SPSS output (f) below under 'PRE_1'. Using this we can calculate odds and odds ratio between 2 specific scores. For example, suppose my present score is 781 and I'd like to know how much increase in odds if I raise my score by 11 points and get 792. Then the odds ratio can be obtained easily. The calculation ends up to an increase of 179% in odds when I raise up my score by 11 points (**Table 4**).

$$\frac{\text{odds at } 792}{\text{odds at } 781} = \frac{1.18}{0.43} = 2.79$$

REFERENCES

1. Allison PD. In: Logistic regression using SAS: theory and application. 2nd ed. Cary (NC, USA): SAS Institute Inc.; 2012. pp. 19-26.

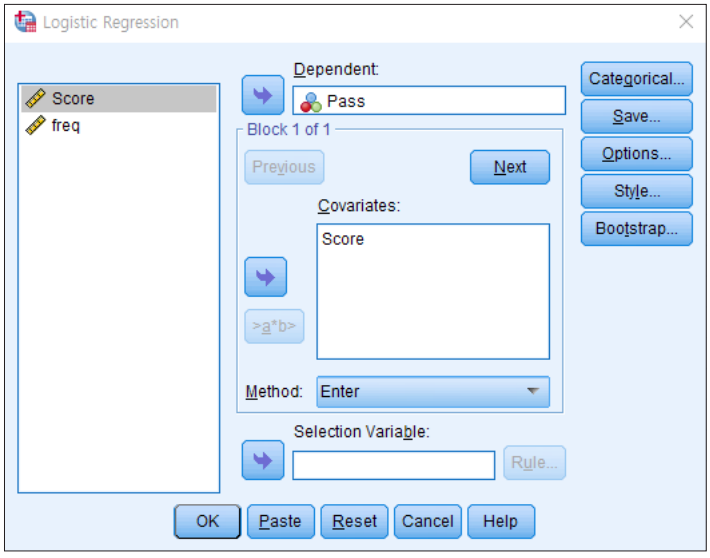
Appendix 1. Procedure of logistic regression using IBM SPSS.

The procedure of logistic regression using IBM SPSS Statistics for Windows Version 23.0 (IBM Corp.) is as follows.

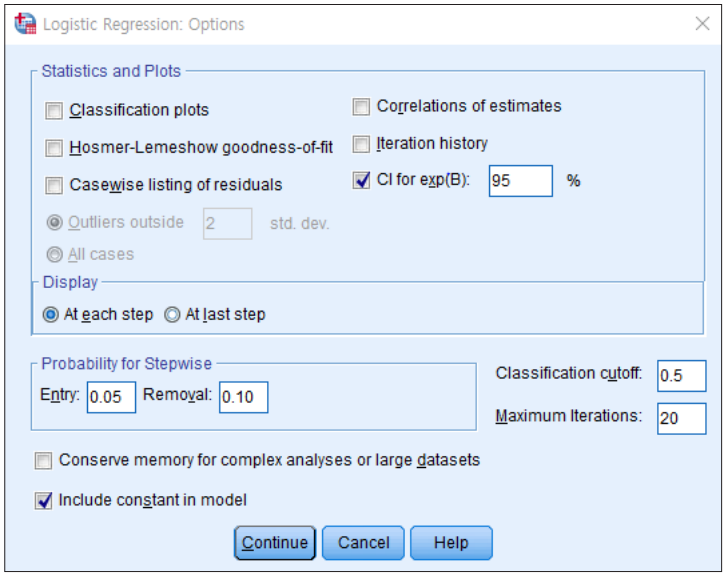
(A) Data (weighted by freq*)

Score	Pass	freq
755	0	100
755	0	100
763	0	100
781	0	100
783	1	100
788	1	100
792	0	100
793	1	100
798	0	100
799	0	100
799	1	100
802	1	100
813	1	100
824	1	100
845	1	100

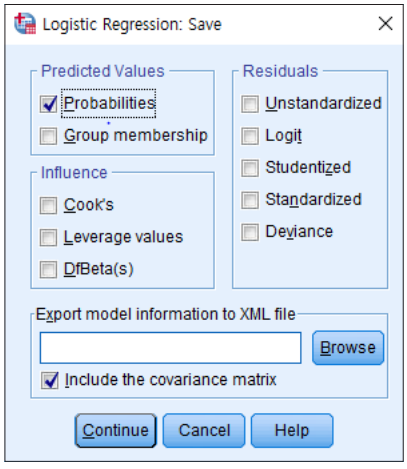
(B) Analyze-Regression-Binary Logistic



(C) Options



(D) Save



(E) Test of model

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	700.826	1	.000
	Block	700.826	1	.000
	Model	700.826	1	.000

(F) Model fit

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1371.944[a]	.373	.498

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

https://doi.org/10.5395/rde.2017.42.4.342

(G) Coefficients

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Score	.093	.005	301.595	1	.000	1.098	1.086	1.109
	Constant	-73.578	4.247	300.105	1	.000	.000		

a. Variable(s) entered on step 1: Score.

(H) Predicted probability

Score	Pass	freq	PRE_1
755	0	100	.03640
755	0	100	.03640
763	0	100	.07370
781	0	100	.29836
783	1	100	.33875
788	1	100	.44935
792	0	100	.54219
793	1	100	.56520
798	0	100	.67434
799	0	100	.69444
799	1	100	.69444
802	1	100	.75032
813	1	100	.89327
824	1	100	.95886
845	1	100	.99397

*In this fictitious data, the 'freq' variable was used to multiply the number of observations to get sufficient power.

https://doi.org/10.5395/rde.2017.42.4.342