일반화선형모형

Logistic Regression

Logistic Regression 모형은 사후확률의 로짓 변환 즉, $\mathbf{logit}(P) = \ln(P/(1-P))$ 이 입력변수의 선형함수로 구성되어 있다고 가정

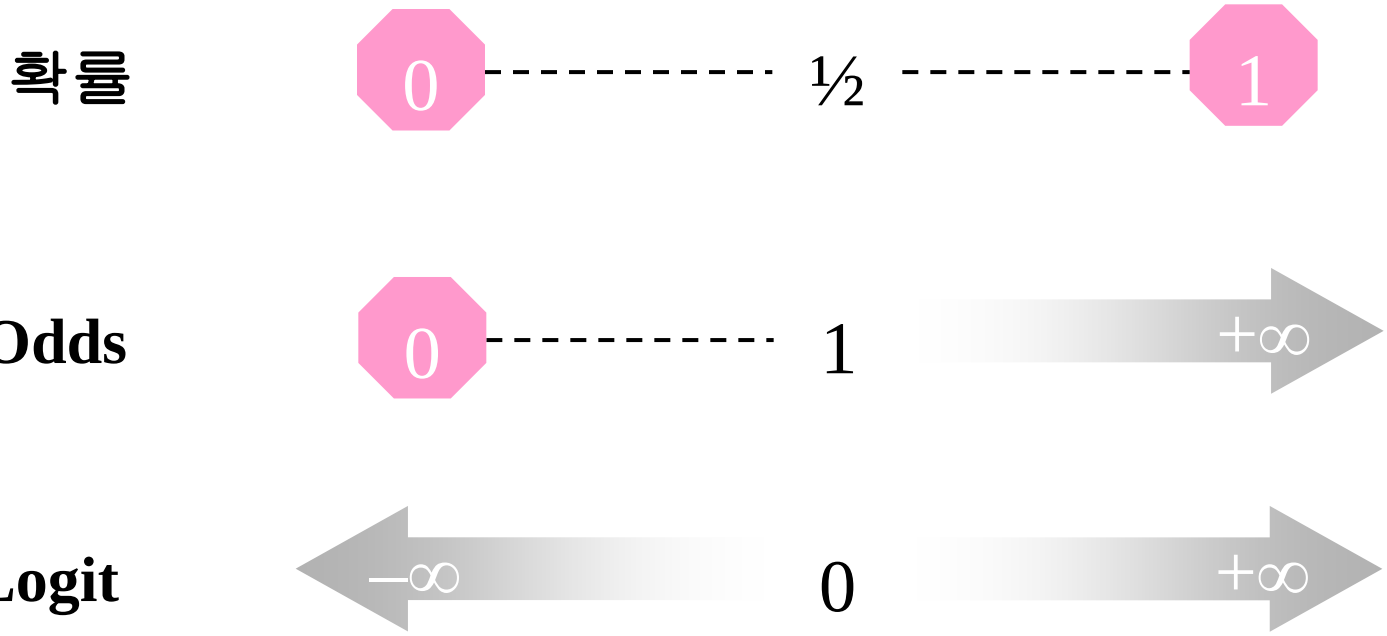
$$\mathbf{logit}(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

$$\mathbf{logit}(P) = \ln(P / (1- P))$$

$x_i = i$ 번째 입력변수

Logistic

$$\text{logit}(P) = \log \text{odds} = \ln\left(\frac{P}{1-P}\right)$$



Multiple Logistic

$$\text{logit}(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

단, $\text{logit}(P) = \ln(P / (1 - P))$

x_i = i 번째 입력변수

$Y \sim$ Bernoulli(p)
<=> 반응변수가 "0" (실패, failure) 또는 "1" (성공, success)의 값을 갖는 베르누이 확률변수

$$0 \le \Pr[Y = 1 \mid \mathbf{x}] = \mu(\mathbf{x}) \le 1$$

$$\text{link function} : \text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \eta(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$$

$$\hat{\mu}(\mathbf{x}) = \text{logit}^{-1}(\hat{\eta}(\mathbf{x}))$$

$$= \text{logistic}(\hat{\eta}(\mathbf{x})) = \frac{1}{1+e^{-\hat{\eta}(\mathbf{x})}} = \frac{1}{1+e^{-\mathbf{x}'\hat{\boldsymbol{\beta}}}}$$

$$= \frac{e^{\hat{\eta}(\mathbf{x})}}{1+e^{\hat{\eta}(\mathbf{x})}} = \frac{e^{\mathbf{x}'\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}'\hat{\boldsymbol{\beta}}}}$$

The **logit** of a number p between 0 and 1 is given by the formula:

$$\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p) = -\log\left(\frac{1}{p} - 1\right).$$

The base of the logarithm function used is of little importance in the present article, as long as it is greater than 1, but the natural logarithm with base e is the one most often used. The choice of base corresponds to the choice of logarithmic unit for the value: base 2 corresponds to a shannon, base e to a nat, and base 10 to a hartley; these units are particularly used in information-theoretic interpretations. For each choice of base, the logit function takes values between negative and positive infinity.

The "logistic" function of any number α is given by the inverse-logit:

$$\operatorname{logit}^{-1}(\alpha) = \operatorname{logistic}(\alpha) = \frac{1}{1+\exp(-\alpha)} = \frac{\exp(\alpha)}{\exp(\alpha)+1}$$

If p is a probability, then $p/(1-p)$ is the corresponding odds; the logit of the probability is the logarithm of the odds. Similarly, the difference between the logits of two probabilities is the logarithm of the odds ratio (R), thus providing a shorthand for writing the correct combination of odds ratios only by adding and subtracting:

$$\log(R) = \log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right) = \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = \operatorname{logit}(p_1) - \operatorname{logit}(p_2).$$

https://en.wikipedia.org/wiki/Logit

The *standard logistic function* is the logistic function with parameters ($k = 1$, $x_0 = 0$, $L = 1$) which yields

$$\begin{aligned} f(x) &= \frac{1}{1 + e^{-x}} \\ &= \frac{e^x}{1 + e^x} \\ &= \tfrac{1}{2} + \tfrac{1}{2} \tanh\left(\tfrac{x}{2}\right) \end{aligned}$$

In practice, due to the nature of the exponential function e^{-x}, it is often sufficient to compute the standard logistic function for x over a small range of real numbers such as a range contained in [−6, +6].

The logistic function has the symmetry property that:

$$1 - f(x) = f(-x).$$

Thus, $x \mapsto f(x) - 1/2$ is an odd function.

The logistic function is an offset and scaled hyperbolic tangent function

$$f(x) = \tfrac{1}{2} + \tfrac{1}{2} \tanh\left(\tfrac{x}{2}\right)$$

or

$$\tanh(x) = 2\, f(2x) - 1\,.$$

This follows from

$$\begin{aligned} \tanh(x) &= \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^x \cdot \left(1 - e^{-2x}\right)}{e^x \cdot \left(1 + e^{-2x}\right)} \\ &= f(2x) - \frac{e^{-2x}}{1 + e^{-2x}} = f(2x) - \frac{e^{-2x} + 1 - 1}{1 + e^{-2x}} = 2f(2x) - 1. \end{aligned}$$

https://en.wikipedia.org/wiki/Logistic_function

Derivative [edit]

The standard logistic function has an easily calculated derivative:

$$f(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$$

$$\frac{d}{dx} f(x) = \frac{e^x \cdot (1+e^x) - e^x \cdot e^x}{(1+e^x)^2}$$

$$\frac{d}{dx} f(x) = \frac{e^x}{(1+e^x)^2} = f(x)(1-f(x))$$

The derivative of the logistic function is an even function:

namely, $f'(-x) = f'(x)$

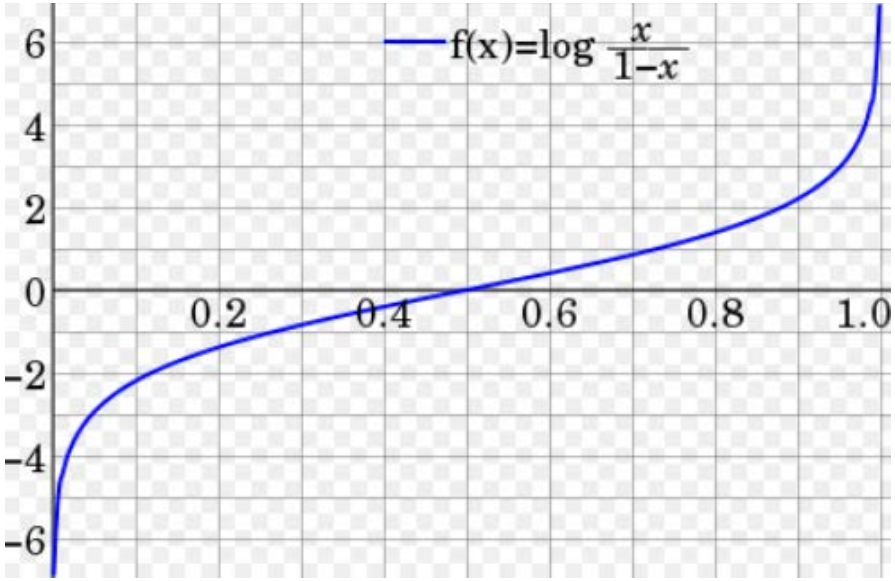
Integral [edit]

Conversely, its antiderivative can be computed by the substitution $u = 1 + e^x$, since $f(x) = \frac{e^x}{1+e^x} = \frac{u'}{u}$, so (dropping the constant of integration):

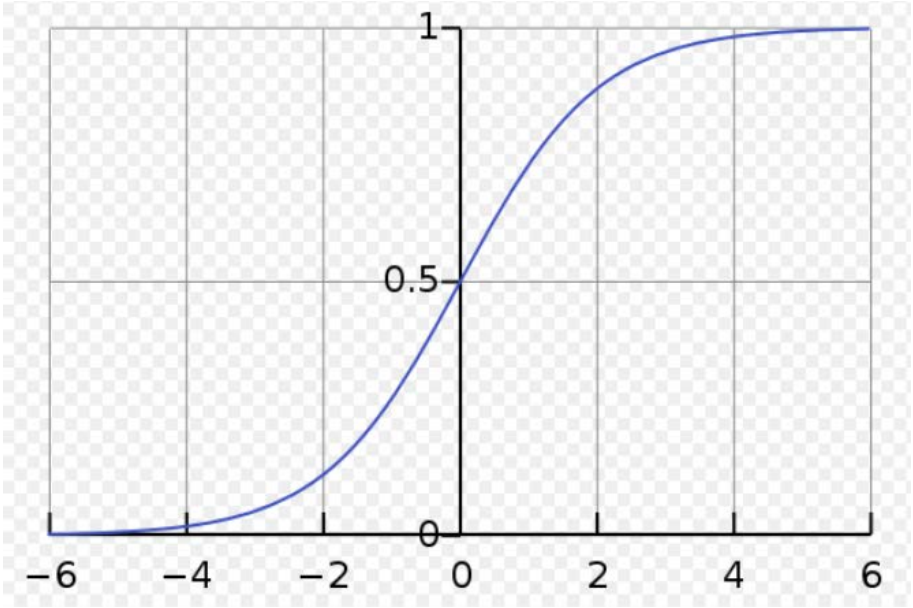
$$\int \frac{e^x}{1+e^x}\,dx = \int \frac{1}{u}\,du = \log u = \log(1+e^x)$$

In artificial neural networks, this is known as the *softplus* function, and (with scaling) is a smooth approximation of the ramp function, just as the logistic function (with scaling) is a smooth approximation of the Heaviside step function.

https://en.wikipedia.org/wiki/Logistic_function



logit function



logistic function

Understanding Logistic Regression

Logistic regression is similar to linear regression except that the target variable is a binary. For example, an Ethologist might assign a crab to one of two classes (say female or male). A male might be coded as $y = 1$, and a female coded by $y = 0$.

Recall, in linear regression the target variable is related to the features via the linear relationship:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{7.1}$$

Suppose $p(y)$ is the probability that a crab is male (we could write $p(y = 1)$ but stick to the shorter notation).

To relate $p(y)$ to the features you might consider writing:

$$p(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{7.2}$$

Unfortunately, this specification, can generate values for $p(y)$ from -∞ to ∞. We need a model that generates probabilities in the 0 to 1 range. This is not guaranteed to be the case if we use equation 7.2. Furthermore, linear regression assumes the values of y are normally distributed. In logistic regression y takes the values 0 or 1, so this assumption is clearly violated.

We need a more appropriate transformation. This can be achieved using the logistic regression model:

$$\ln\left(\frac{p(y)}{1 - p(y)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{7.3}$$

It is a non-linear transformation of the linear regression model where the target variable is binary.

Odds Ratio

When we have a binary classification problem, we are typically interested in the probability that an observation belongs to a specific class. Statisticians often think of this in terms of odds.

For example, in a sample of crabs, the odds of being a male can be calculated as:

$$Odds = \frac{Number\,of\,Male\,Crabs}{Total\,Number\,of\,Crabs}$$

or, equivalently:

$$Odds = \frac{p(y)}{1 - p(y)}$$

The ratio $\left(\frac{p(y)}{1-p(y)}\right)$ is called the odds ratio.

Log odds ratio

The natural logarithm of the odds ratio is called the log odds ratio or logit. Let's investigate what an odds ratio of 1 implies:

$$\ln\left(\frac{p(y)}{1 - p(y)}\right) = 1$$

$$\Rightarrow p(y) = [1 - p(y)]$$

$$\Rightarrow p(y) + p(y) = 1$$

$$\Rightarrow p(y) = 0.5$$

In other words, an odds ratio equal to 1 implies the probability that $y = 1$ is 0.5.

Understanding Logistic Regression

Logistic regression is similar to linear regression except that the target variable is a binary. For example, an Ethologist might assign a crab to one of two classes (say female or male). A male might be coded as $y = 1$, and a female coded by $y = 0$.

Recall, in linear regression the target variable is related to the features via the linear relationship:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{7.1}$$

Suppose $p(y)$ is the probability that a crab is male (we could write $p(y = 1)$ but stick to the shorter notation).

To relate $p(y)$ to the features you might consider writing:

$$p(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{7.2}$$

Unfortunately, this specification, can generate values for $p(y)$ from -∞ to ∞. We need a model that generates probabilities in the 0 to 1 range. This is not guaranteed to be the case if we use equation 7.2. Furthermore, linear regression assumes the values of y are normally distributed. In logistic regression y takes the values 0 or 1, so this assumption is clearly violated.

We need a more appropriate transformation. This can be achieved using the logistic regression model:

$$\ln\left(\frac{p(y)}{1 - p(y)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{7.3}$$

It is a non-linear transformation of the linear regression model where the target variable is binary.

Since p is a probability, we can see that the logistic regression model is constructed so that $0 \leq p \leq 1$. Indeed, from equation 7.3 as:

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ...\beta_k x_k$$

becomes very large, p approaches 1; and as:

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ...\beta_k x_k$$

becomes very small, p approaches 0; Furthermore, if:

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ...\beta_k x_k = 0$$

then $p = 0.5$.

Interpreting coefficients

We interpret $exp(\beta_i)$ as the effect of the independent variables or features on the odds ratio. For example, if we postulate the logistic regression:

$$\ln\left(\frac{p(y)}{1-p(y)}\right) = \alpha + \beta_1 x_1$$

and on estimation find that $\hat{\beta} = 0.963$, so that $exp(\hat{\beta}) = 1.999$. This implies a 1 unit change in x_1 would make the event $y = 1$ about twice as likely.

> ***NOTE...*** ✍
>
> In statistics textbooks you will often see $\ln\left(\frac{p(y)}{1-p(y)}\right)$ called the logit transform or simply logit.

The Logistic Curve

The logistic curve or sigmoid function, captures the relationship between a binary target variable and features. It is calculated as:

$$p(y) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{7.4}$$

Because the relationship between $p(y)$ and x is nonlinear, the parameters α and β do not have a straightforward interpretation as they do in linear regression.

Figure 7.1 shows the logistic curve. It is bounded by 0 and 1, and therefore can be interpreted in terms of probabilities. The curve is symmetric about the point where $x = -\frac{\alpha}{\beta}$. In fact, the value of $p(y)$ is 0.5 for this value of x.

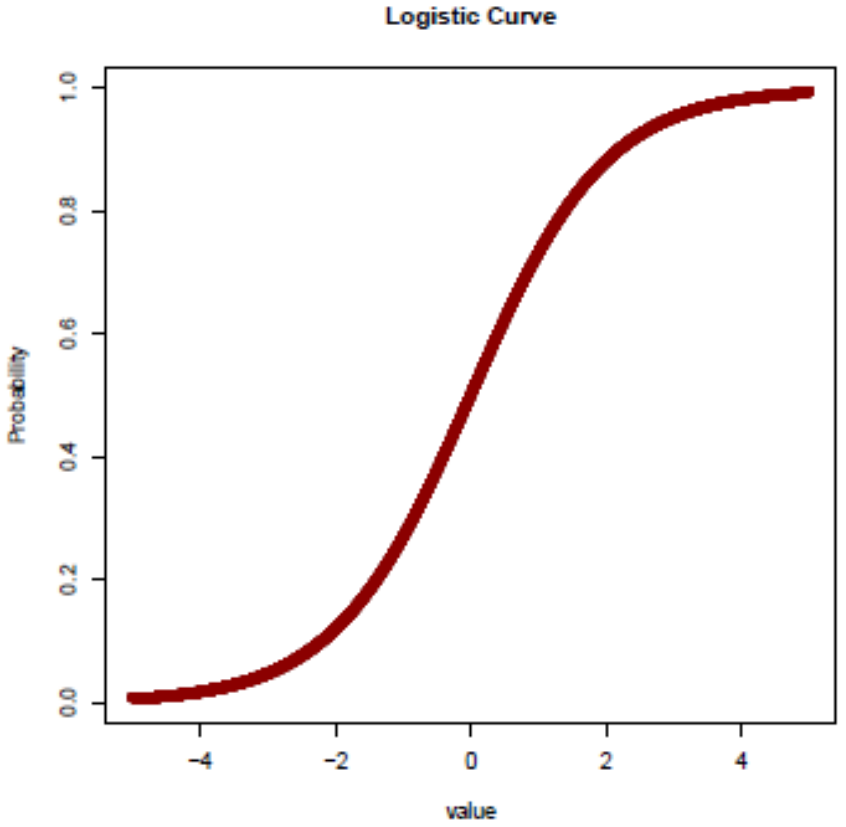


Figure 7.1: Logistic Curve

> ***NOTE...*** ✍
>
> The values of α and β, determine the location and spread of the logistic curve.

Relationship to logistic regression

If we were to run a logistic regression on the feature x we would specify:

$$\log\left(\frac{p(y)}{1-p(y)}\right) = \alpha + \beta x$$

In terms of the odds we can rewrite the above as:

$$\frac{p(y)}{1-p(y)} = \exp(\alpha + \beta x)$$

Of course, our interest is in $p(y)$; it is given by:

$$p(y) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)},$$

which is equation 7.4 we saw earlier.
The key points to note are that:

1. The log odds are linearly related to the features.
2. The relation between the features and $p(y)$ is nonlinear taking the S-shaped curve of Figure 7.1.

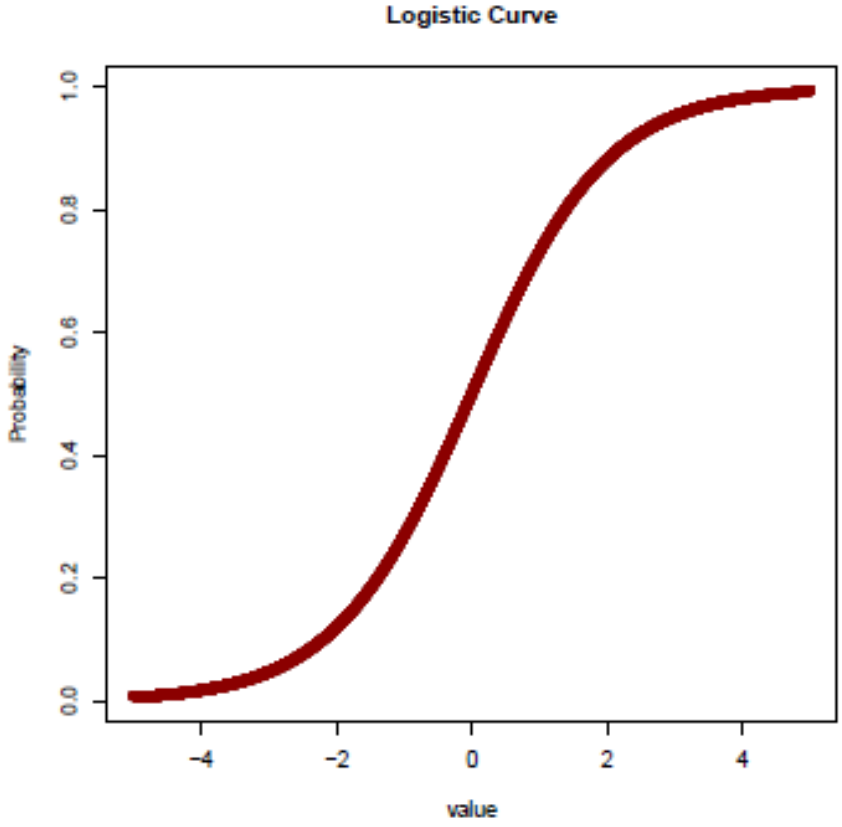


Figure 7.1: Logistic Curve

NOTE... ✍

Similar to linear regression, logistic regression assumes the features are independent. However, it does not make any assumptions about the probability distribution of the features.

NOTE... ✍

The values of α and β, determine the location and spread of the logistic curve.

Maximum Likelihood Estimation

In linear regression, the method of ordinary least squares can be used to estimate the regression coefficients. In logistic regression, we use a different approach called maximum likelihood estimation (MLE). MLE is a very general approach to obtain estimates of the parameters of probability models. Similar to ordinary least squares, the goal is to find the smallest possible deviance between the observed (y) and predicted values ($\hat{y}$).

The idea behind MLE is to choose the most likely values of the parameters α and β given, the observed sample, say $\{x_1, ...x_n\}$. Intuitively, the actual values these parameters take should depend in some way on the values observed in the sample data. This link is established via a probability model, which we denote by $f(x)$. The probability model is used to form the likelihood equation.

Probability model & likelihood equation

In logistic regression, the probability model is based on the binomial distribution, where:

$$f(x,p) = \begin{cases} \theta & if\ y_i = 1 \\ 1-\theta & if\ y_i = 0 \end{cases}$$

In other words, the probability of x_i being a male crab ($y_i = 1$) occurs with probability θ. And therefore:

$$p(y_i = 1) = \theta = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

.

The likelihood equation is given by:

$$L = \prod_{i=1}^{n} \theta_i^{y_1} (1-\theta_i)^{1-y_1}$$

Statisticians figured out a while back that is easier to work in natural logarithms. Taking the natural log of both sides and simplifying we get the log likelihood equation:

$$LL = \ln\ L = \sum_{i=1}^{n} [y_i \ln \theta_i + (1-y_i)\ln(1-\theta_i)]$$

Whilst ordinary least squares, in linear regression, minimizes the residual sum of squares; MLE in logistic regression minimizes model deviance. The MLE estimate of α and β, that minimize model deviance, are retrieved as the values $\hat{\alpha}$ and $\hat{\beta}$. These values maximize the probability of the observed data given the specified probability model.

NOTE... ✍

The maximum likelihood estimate of the parameters α and β are the values that maximizes the probability of the sample data $\{x_1, ...x_n\}$.

As a Bayes Classifier

A direct link to the Bayes classifier can be observed if we specify the logistic regression formula as a probability distribution of the class posterior probabilities. For a two class classification problem we have:

$$P(c_i = 1|x) = \frac{1}{[1 + \exp[-\alpha - \sum_{i=1}^{n} \beta_i x_i]]}$$

$$P(c_i = 0|x) = \frac{\exp[-\alpha - \sum_{i=1}^{n} \beta_i x_i]}{[1 + \exp[\alpha + \sum_{i=1}^{n} \beta_i x_i]]}$$

In this case we predict $c_i = 1$ if:

$$P(c_i = 1|x) > P(c_i = 0|x),$$

which is essentially the decision criteria used for the Bayes classifier.

로지스틱모형

deviance: 모형의 적합도 (goodness of fit)를 나타내는 측도
선형모형에서의 잔차의 제곱합을 일반화한 측도
최대가능도법에 의하여 구해진 추정값에 대하여 모형의 적합도를 가능도함수로 나타낸 측도

<R>
Null Deviance = 2[LL(saturated model) - LL(null model)] on df = df_sat - df_null
모형을 적합하기 전의 deviance
Residual Deviance = 2[LL(saturated model) - LL(proposed model)] on df = df_sat - df_prop
모형을 적합 한 후의 deviance

LL: 각 모형에서 최대화된 로그가능도함수

귀무가설 (모형이 적합하지 않음) 하에서 두 deviance의 차이는 근사적으로 카이제곱 분포를 이룸

로지스틱모형

이항편차 (binomial deviance) :

$$D = 2\sum_{i=1}^{n}\left[y_i \log(y_i / \hat{\mu}_i) + (1-y_i)\log((1-y_i)/(1-\hat{\mu}_i))\right]$$

(*i*) all $y_i = \hat{\mu}_i \Rightarrow D = 0$

(*ii*) all $y_i \neq \hat{\mu}_i \Rightarrow D = \infty$

로지스틱모형

AIC (Akaike Information Criterion)

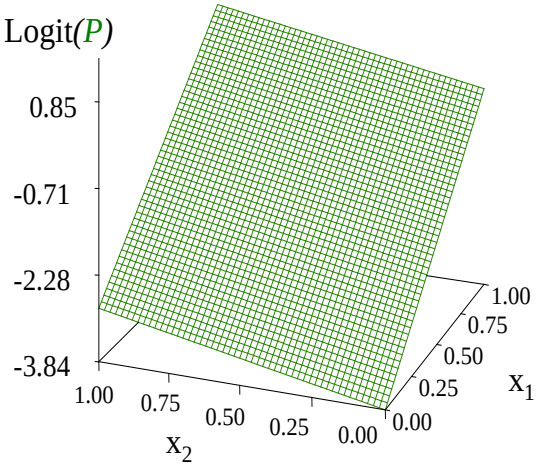
$$AIC = 2k - 2\ln(L)$$

k: 모형모수의 개수, 복잡도
L: 주어진 모형으로 최대화한 가능도

*AIC*값이 작은 모형을 찾음
모형이 복잡해지면 $-\ln(L)$은 작아지나 *k*가 커짐
모형이 간단해지면 k가 작아지나 $-\ln(L)$이 커짐

Example

$$\text{logit}(P) = -3.84 + 3.73 \cdot x_1 + 0.966 \cdot x_2$$



- x_1이 **1**단위 증가함에 따라 사건이 발생하지 않을 확률에 대하여 사건이 발생할 확률의 비율, 즉 **Odds** 비의 자연 로그 값이 **3.73**씩 증가함을 의미

- 추정된 로지스틱 모형은 다음과 같이 확률척도로 역 변환할 수 있다.

$$P = \frac{1}{1 + \exp\left(-\left(-3.84 + 3.73\,x_1 + 0.966\,x_2\right)\right)}$$

변수선택의 기준

- *adjusted* $R^2 = 1 - \frac{n-1}{n-p-1}\frac{SSE_p}{SST}$

SSE_p : *SSE of reduced* model

(*no. of* exp. var.: p)

- *Mallow's* C_p : $\frac{SSE}{\hat{\sigma}^2} + 2(p+1) - n$
- *Akaike Information Criterion*(AIC):

$$AIC = -2\log(ML) + 2(p+1) = n\log\frac{SSE}{n} + 2(p+1)$$

- *Bayesian Information Criterion*(BIC):

$$BIC = n\log\frac{SSE}{n} + 2(p+3)q - 2q^2$$

$$q = \frac{\hat{\sigma}^2}{SSE/n}$$

Advantages of Logistic Regression

- Since the log odds takes the values between $-\infty$ to ∞, it allows the properties of linear regression to be exploited via a nonlinear relationship between the target variable and features.
- In addition, probabilities (or odds ratios) can be directly calculated from the model parameters.
- Unlike linear discriminant analysis, the features are not assumed to be normally distributed, or have equal variance in each class.
- This makes it more robust.

Limitations of Logistic Regression

- Logistic regression is a discriminative classifier.
- However, in practice, generative classifiers such as naive Bayes often outperform discriminative classifiers such as logistic regression.
- Well over a decade ago scholars Andrew Ng and Michael Jordan studied the error properties of logistic regression and naive Bayes models.
- They found that naive Bayes reaches its asymptotic error bound much faster than the discriminative logistic regression classifier.
- However, as the sample size grows larger, and larger, logistic regression outperforms the naive Bayes classifier.
- The scholars illustrated the result for 15 real-world data-sets.
- It turns out the generative naive Bayes model reaches its asymptotic bound at a rate $O(logN)$, whilst the discriminative logistic model approaches it bound at a rate of $O(N)$.
- The practical implication of this is that the naive Bayes model reaches its asymptotic solution for a much smaller data sample than the logistic model.
- In other words, if you have a large sample try logistic regression.
- If you have a small sample try naïve Bayes.

Example 1
Logistic Regression

To illustrate how information can be gleaned from the plot, consider the Temperature and O-ring Failure dataset (Myers *et al.* [9], p.126) from the NASA space shuttle Challenger accident (January 28, 1986). The space shuttle *Challenger* exploded shortly after the launch and crashed into the Atlantic Ocean off the coast of Florida, killing all seven astronauts aboard. The cause of the accident was later determined to be the failure of the O-rings on the solid rocket booster. The O-rings failed because they lost flexibility at low temperatures (^0F). This dataset with 24 observations consists of two variables: the explanatory variable, x, temperature at launch (^0F) and a categorical response variable, y, (0 = no O-ring failures and 1 = At Least One O-Rings Failure). Table 1 and Figure 1 show the original data and a scatterplot of the dataset, respectively.

We can fit this dataset with the logistic regression model of the form:

$$E(y_i) = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})},$$

where $\boldsymbol{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x$ for this example.

Table 1. Temperature and O-ring failure dataset.

NO	Temperature	Failure	NO	Temperature	Failure
1	53	1	13	70	1
2	56	1	14	70	1
3	57	1	15	72	0
4	63	0	16	73	0
5	66	0	17	75	0
6	67	0	18	75	1
7	67	0	19	76	0
8	67	0	20	76	0
9	68	0	21	78	0
10	69	0	22	79	0
11	70	0	23	80	0
12	70	1	24	81	0

(Temperature: temperature at launch (^{0}F), Failure: at least one O-ring failure)

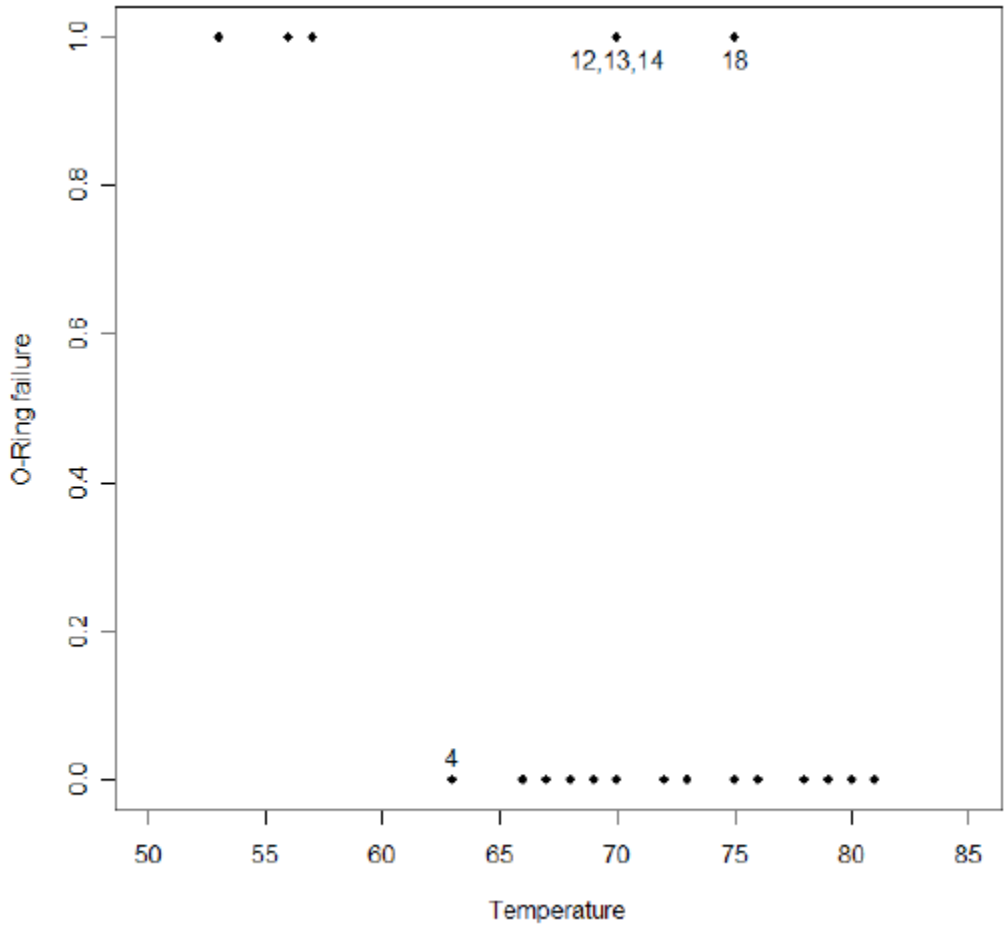


Figure 1. Scatterplot for temperature and O-ring failure dataset.

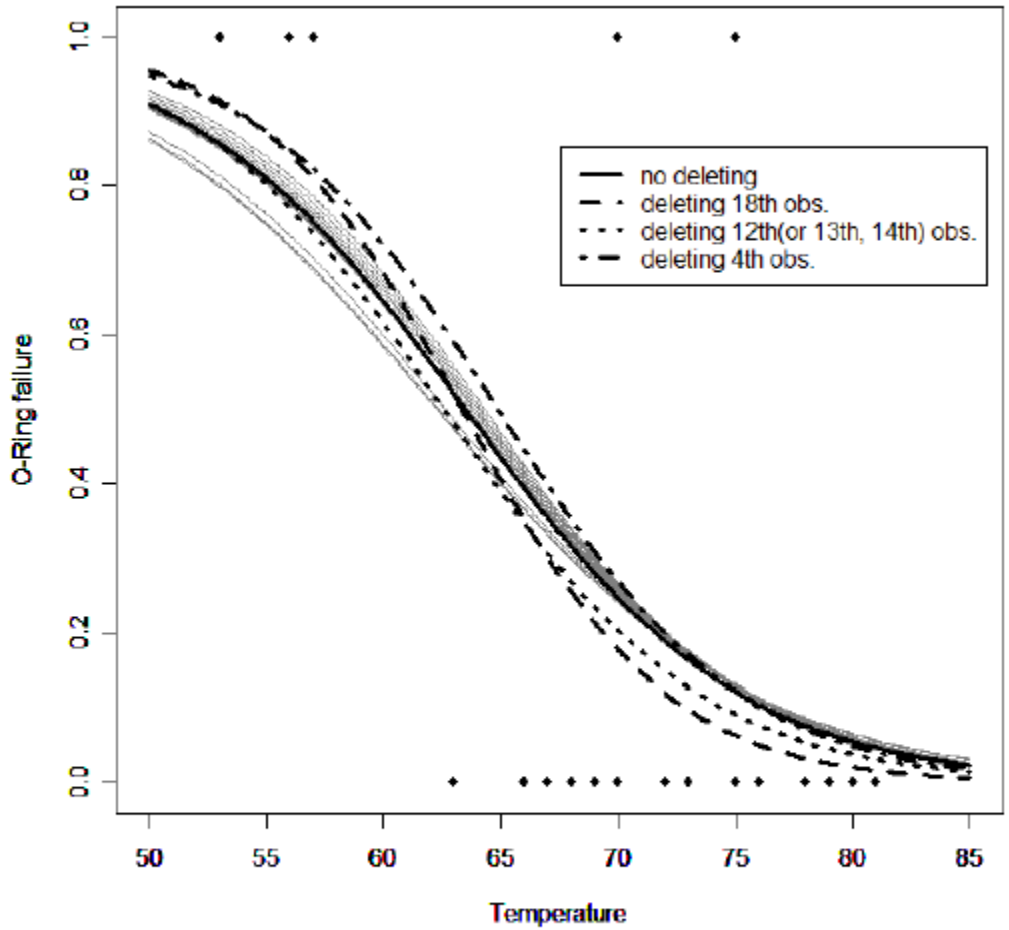


Figure 3. Logistic regression curves with respect to 24 cases in which each single observation is deleted respectively.

Example 2
Logistic Regression

The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix (See Faraway [4]) to examine patterns in their overall health to determine indicators of diabetes. This dataset contains 9 variables (pregnant (number of times pregnant), glucose (plasma glucose concentration at 2 hours in an oral glucose tolerance test), diastolic (diastolic blood pressure (mm Hg)), triceps (triceps skin fold thickness (mm)), insulin (2-Hour serum insulin (mu U/ml)), bmi (body mass index (weight in kg/(height in meters squared))), diabetes (diabetes pedigree function), age (in years)), diabetes test result (coded 0 if negative, 1 if positive)).

We fit a logistic regression model with the result of the diabetes test as the response and all the other variables as predictors. As the result of the test for the significance of main effects, we obtain a final model with 5 significant variables (x_1 : pregnant, x_2 : glucose, x_3 : diastolic, x_4 : bmi, x_5 : diabetes) based on the 768 observations as

$$\hat{y} = \frac{\exp(\mathbf{x}'\mathbf{b})}{1+\exp(\mathbf{x}'\mathbf{b})},$$

where $\mathbf{x}'\mathbf{b} = -7.9550 + 0.1535x_1 + 0.0347x_2 - 0.0120x_3 + 0.0848x_4 + 0.9106x_5$.

Example 3
Poisson Regression

For GLMs, we can also use the pairwise firework plot matrix for detecting influential points when the number of parameters or summaries grows beyond what can easily be visualized with the 3-D plot. In the pairwise firework plot matrix, each panel uses curves to connect the values of two of the weighted least-squares estimators or weighted deviance calculated by changing weights for each observation from 1 to 0. With this firework plot matrix, we can identify and quantify the impact of outliers and influential points on the estimated parameters in generalized linear model analysis. As the dimension of the input factor space increases it becomes more difficult to identify outliers and influential points because proximity of points to each other becomes less apparent and more difficult to visualize. In addition for GLMs, the relationship of the inputs and the response is less direct than with standard linear models. To illustrate how information can be gleaned from the plot, we consider the wave solder dataset (Myers *et al.* [8], p.459), which models the number of defects in the solder joint for an electronic circuit card assembly. This dataset consists of a response variable and seven factors (y: the number of defects in the solder joint, A: prebake condition, B: flux density, C: conveyor speed, D: preheat condition, E: cooling time, F: ultrasonic solder agitator, G: solder temperature) with 47 observations. Each factor is at two levels, and the experimental design is a replicated 2^{7-3} fractional factorial design. The complete data set is provided in Appendix A.

Appendix A: Wave solder dataset

NO	A	B	C	D	E	F	G	*y*
1	-1	-1	-1	-1	-1	-1	-1	13
2	-1	-1	-1	-1	-1	-1	-1	30
3	-1	-1	-1	-1	-1	-1	-1	26
4	-1	-1	-1	1	1	1	1	4
5	-1	-1	-1	1	1	1	1	16
6	-1	-1	-1	1	1	1	1	11
7	-1	-1	1	-1	-1	1	1	20
8	-1	-1	1	-1	-1	1	1	15
9	-1	-1	1	-1	-1	1	1	20
10	-1	-1	1	1	1	-1	-1	42
11	-1	-1	1	1	1	-1	-1	43
12	-1	-1	1	1	1	-1	-1	64
13	-1	1	-1	-1	1	-1	1	14
14	-1	1	-1	-1	1	-1	1	15
15	-1	1	-1	-1	1	-1	1	17
16	-1	1	-1	1	-1	1	1	10
17	-1	1	-1	1	-1	1	-1	17
18	-1	1	-1	1	-1	1	-1	16
19	-1	1	1	-1	1	1	-1	36
20	-1	1	1	-1	1	1	-1	29
21	-1	1	1	-1	1	1	-1	53
22	-1	1	1	1	1	1	1	5
23	-1	1	1	1	1	1	1	9
24	-1	1	1	1	-1	-1	1	16
25	1	-1	-1	-1	1	1	-1	29
26	1	-1	-1	-1	1	1	-1	0
27	1	-1	-1	-1	1	1	-1	14
28	1	-1	-1	1	-1	-1	1	10
29	1	-1	-1	1	-1	-1	1	26
30	1	-1	-1	1	-1	-1	1	9
31	1	-1	1	-1	1	1	1	28
32	1	-1	1	-1	1	1	1	19
33	1	-1	1	1	-1	1	-1	100
34	1	-1	1	1	-1	1	-1	129
35	1	-1	1	1	-1	1	-1	151
36	1	1	-1	-1	-1	1	1	11
37	1	1	-1	-1	-1	1	1	15
38	1	1	-1	-1	-1	1	1	11
39	1	1	-1	1	1	-1	-1	17
40	1	1	-1	1	1	-1	-1	2
41	1	1	-1	1	1	-1	-1	17
42	1	1	1	-1	1	-1	-1	53
43	1	1	1	-1	1	-1	-1	70
44	1	1	1	-1	1	-1	-1	89
45	1	1	1	1	1	1	1	23
46	1	1	1	1	1	1	1	22
47	1	1	1	1	1	1	1	7

We can fit this dataset with a Poisson regression model as follows

$$E(y_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta}) .$$

After examining all of the AIC values for model selection, the best model with 47 observations is selected to be

$$\hat{\mu} = \exp(3.0769 + 0.4405C - 0.4030G + 0.2821AC - 0.3113BD) ,$$

with the minimum AIC=475.85 and deviance value 241.78.

R library

- lm(): linear model
- glm(): generalized linear model
- gam(): generalized additive model
- lme(): linear mixed-effects model
- lmer(): linear mixed-effects model
- nls(): non-linear model

불균형 자료 (unbalanced data)의 분석

- 현실적인 분류 문제에서는 두 그룹의 자료의 수에 큰 차이가 나는 경우가 많음 (불균형 자료)
- 예: 사기 방지를 위한 분류문제에서 정상자료의 수가 전체 자료의 99% 이상
- 원 자료를 그대로 사용하지 않고 정상자료에서 적절한 수의 표본을 사용하여 모형을 구축하는 방법이 실제 자료분석에서 널리 사용됨
- 자료의 크기가 매우 커서 양을 줄이고자 할 때 이러한 표본추출법이 매우 유용함
- 주의할 점: 모형의 해석
- 이유: 사기 방지를 위한 분류문제에서 정상자료에서 사기자료 수 만큼의 자료를 추출하여 사용할 경우 원 자료에서 사기자료의 비율은 매우 작은 반면 추출된 자료에서 사기자료의 비율은 50%이기 때문임

불균형 자료 (unbalanced data)의 분석

- 불균형 자료에서 그룹별로 다른 확률로 자료를 추출하는 방법은 임상분야에서 많이 사용되고 있는 사례-대조 연구 (case-control study)와 같은 방법임
- 이 경우 로지스틱 회귀모형에서 y-절편항 beta_0는 편의추정량, beta_k는 점근적 일치추정량이 됨
- 추출된 자료를 이용하여 로지스틱 회귀모형을 적합하는 경우 확률의 추정은 불가능하나 자료들 사이의 확률의 대소는 추정 가능함
- Pr[Y=1|x]는 알 수 없으나 주어진 두 개의 입력변수 x_1과 x_2에 대하여 Pr[Y=1|x_1] > Pr[Y=1|x_2]의 여부는 알 수 없음
- 이러한 이유로 실제 자료분석에서 로지스틱 회귀모형으로부터 추정된 확률을 점수 (score)라고 부름
- 원 자료에서 두 그룹의 확률 Pr[Y=1] 과 Pr[Y=0]이 알려져 있는 경우 추출된 자료로부터 추정된 확률에 대한 적절한 변환을 통하여 원 자료에서의 확률을 추정할 수 있음
- 이러한 작업을 보정 (calibration)이라고 함
- 마케팅에서 우수고객의 추출에 관심이 있는 경우 점수의 추정으로 충분하지만 위험도분석 (risk analysis)에서는 확률 자체에 관심이 있으므로 확률 추정값의 보정이 필수적임