

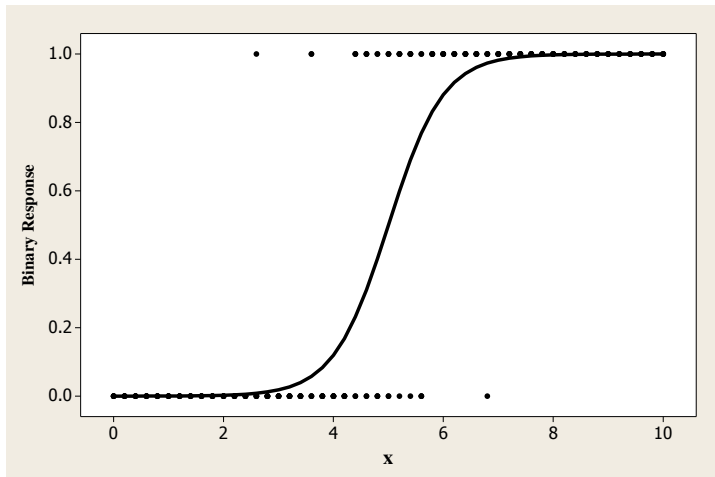
Logistic Regression

Logistic Regression

A simple binary logistic regression with one continuous predictor variable looks something like:

Logistic Regression

A simple binary logistic regression with one continuous predictor variable looks something like:



Theory Behind Logistic Regression

For logistic regression, the responses Y_i are 0 and 1 where 1 represents some event occurring (such as survival, death, cancer, etc...) and 0 represents that event not occurring. You want to predict Y_i from some predictor variables measured on individual i : $X_{1,i}, X_{2,i}, \dots, X_{k,i}$.

Theory Behind Logistic Regression

For logistic regression, the responses Y_i are 0 and 1 where 1 represents some event occurring (such as survival, death, cancer, etc...) and 0 represents that event not occurring. You want to predict Y_i from some predictor variables measured on individual i : $X_{1,i}, X_{2,i}, \dots, X_{k,i}$.

It isn't reasonable to predict only 0's and 1's. What is reasonable, is to predict the probability of the event for an individual. Let π denote the probability we are modeling. We model the log odds as being linear:

Theory Behind Logistic Regression

For logistic regression, the responses Y_i are 0 and 1 where 1 represents some event occurring (such as survival, death, cancer, etc...) and 0 represents that event not occurring. You want to predict Y_i from some predictor variables measured on individual i : $X_{1,i}, X_{2,i}, \dots, X_{k,i}$.

It isn't reasonable to predict only 0's and 1's. What is reasonable, is to predict the probability of the event for an individual. Let π denote the probability we are modeling. We model the log odds as being linear:

$$\text{logit}(\pi) = \log \frac{\pi}{1 - \pi} = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

Theory Behind Logistic Regression

For logistic regression, the responses Y_i are 0 and 1 where 1 represents some event occurring (such as survival, death, cancer, etc...) and 0 represents that event not occurring. You want to predict Y_i from some predictor variables measured on individual i : $X_{1,i}, X_{2,i}, \dots, X_{k,i}$.

It isn't reasonable to predict only 0's and 1's. What is reasonable, is to predict the probability of the event for an individual. Let π denote the probability we are modeling. We model the log odds as being linear:

$$\text{logit}(\pi) = \log \frac{\pi}{1 - \pi} = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

The probability π can be recovered by

$$\pi = \frac{e^{\text{logit}(\pi)}}{1 + e^{\text{logit}(\pi)}}$$

Theory Behind Logistic Regression

The estimates of the coefficients do not come from least squares but from maximum likelihood:

$$\Pr(\text{data}|\beta_0, \beta_1, \dots, \beta_k) = \prod_{i=1}^n \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i}$$

where the terms in the product are a cheap way of getting π_i if $Y_i = 1$ and $(1 - \pi_i)$ if $Y_i = 0$ and the π_i 's depend on the coefficients through the link function as above.

Theory Behind Logistic Regression

The coefficients that make the above probability the greatest are the maximum likelihood estimators and they are found numerically in statistical software packages such as Minitab. Statisticians like maximum likelihood estimators for the following reasons:

Theory Behind Logistic Regression

The coefficients that make the above probability the greatest are the maximum likelihood estimators and they are found numerically in statistical software packages such as Minitab. Statisticians like maximum likelihood estimators for the following reasons:

1. the estimates are asymptotically unbiased

Theory Behind Logistic Regression

The coefficients that make the above probability the greatest are the maximum likelihood estimators and they are found numerically in statistical software packages such as Minitab. Statisticians like maximum likelihood estimators for the following reasons:

1. the estimates are asymptotically unbiased
2. the estimates are asymptotically efficient (that is, they have the smallest variance possible)

Theory Behind Logistic Regression

The coefficients that make the above probability the greatest are the maximum likelihood estimators and they are found numerically in statistical software packages such as Minitab. Statisticians like maximum likelihood estimators for the following reasons:

1. the estimates are asymptotically unbiased
2. the estimates are asymptotically efficient (that is, they have the smallest variance possible)
3. the estimates are asymptotically normally distributed

Theory Behind Logistic Regression

The problem is that all of these are “asymptotically” - as you collect more and more data, the results hold. But you don't in fact have any guarantee that the estimates are good with a data set of size 10 or of size 100 or even of size 1000 necessarily. All you know is that if you keep collecting data, eventually the properties of unbiasedness, efficiency, and normality will hold.

Theory Behind Logistic Regression

What is the interpretation of the coefficients of a logistic regression? Suppose that you have a model

Theory Behind Logistic Regression

What is the interpretation of the coefficients of a logistic regression? Suppose that you have a model

$$\text{logit}(\pi(x)) = \log \frac{\pi(x)}{1 - \pi(x)} = 1 - 2x.$$

Theory Behind Logistic Regression

What is the interpretation of the coefficients of a logistic regression? Suppose that you have a model

$$\text{logit}(\pi(x)) = \log \frac{\pi(x)}{1 - \pi(x)} = 1 - 2x.$$

The coefficient -2 gives the additive decrease in log odds if x is increased by 1 unit. That is

$$\text{logit}(\pi(x + 1)) - \text{logit}(\pi(x)) = 1 - 2(x + 1) - (1 - 2x) = -2$$

Theory Behind Logistic Regression

What is the interpretation of the coefficients of a logistic regression? Suppose that you have a model

$$\text{logit}(\pi(x)) = \log \frac{\pi(x)}{1 - \pi(x)} = 1 - 2x.$$

The coefficient -2 gives the additive decrease in log odds if x is increased by 1 unit. That is

$$\text{logit}(\pi(x + 1)) - \text{logit}(\pi(x)) = 1 - 2(x + 1) - (1 - 2x) = -2$$

Thus the odds decrease by a factor of e^{-2} :

Theory Behind Logistic Regression

What is the interpretation of the coefficients of a logistic regression? Suppose that you have a model

$$\text{logit}(\pi(x)) = \log \frac{\pi(x)}{1 - \pi(x)} = 1 - 2x.$$

The coefficient -2 gives the additive decrease in log odds if x is increased by 1 unit. That is

$$\text{logit}(\pi(x + 1)) - \text{logit}(\pi(x)) = 1 - 2(x + 1) - (1 - 2x) = -2$$

Thus the odds decrease by a factor of e^{-2} :

$$\frac{\pi(x + 1)}{1 - \pi(x + 1)} = e^{-2} \frac{\pi(x)}{1 - \pi(x)} \approx .135 \frac{\pi(x)}{1 - \pi(x)}$$

Thus the odds of the event decrease by 86.5% for every increase of one unit in x .

Odds Ratios

If the probability of an event is small, then the odds ratio is approximately that probability:

One out of eight women will be diagnosed with breast cancer at some point in her life.

$$\pi = \frac{1}{8} = 0.125 \quad \text{Odds} = \frac{1/8}{(1 - 1/8)} = \frac{1}{7} = 0.143$$

Women who have their first child before age 20 have half the risk of developing breast cancer as women who have their first child after age 30. Using $1/8$ as the probability that a woman who has her first child after age 30 develops breast cancer, we see that the effect on the odds ratio is more dramatic:

$$\text{Odds Ratio} = \frac{\frac{1/16}{1-1/16}}{\frac{1/8}{1-1/8}} = \frac{1}{2} \left(\frac{1 - 1/8}{1 - 1/16} \right) = 0.467$$