Logistic regression

Dichotomous response

N. Torelli, G. Di Credico, V. Gioia

2024

University of Trieste

Introduction

GLM: introduction and basic ideas

- GLMs allow to extend classical normal linear models in many directions:
  - response variables can be assumed non-normal (*including discrete distributions or distributions with support* $[0, \infty)$);
  - The mean and the variance of the response are assumed to vary according to values of observed covariates
  - The impact of covariates on the mean of the response is specified according to a (possibly) *non-linear* function of a linear combination of the covariates
- Main advantages are:
  - Unification of seemingly different models: it makes easy to use, understand and teach the techniques. Many of the standard ways of thinking LM carry over to GLMs;
  - Normal LMs, probit and logit models, log-linear models for contingency tables, Poisson regression, some survival analysis models are GLMs;
  - A single general theory and a single general computational algorithm can be developed for inference.

Regression for dichotomous response: Logistic regression

Dichotomous response: Some examples

- In many cases the variable of interest is not a quantitative (numeric) variable.
- The simplest, yet interesting, case is the one where the response variable is dichotomous. Very often we observe for a sample of units whether an event occurred or not. Examples of applications could be:
    - whether a person prefer to use an electric vehicle
    - whether a person purchases an item
    - whether a person decides to to change Adsl provider
    - whether a individual decides to retire o to continue to work in a given year
    - whether a firm becomes insolvent
    - whether a individual has a defined disease

Binary dependent variable

- As in the case of quantitative response variables we are interested in building a statistical model that allows us to predict whether a specific event occurrs (or if a unit belongs to one class).
- Exactly like in standard linear regression model we aim to explain (predict) a dependent variable y_i by using observed characteristics of the i–th unit such as their age, sex, education, income, etc..
- A dichotomous dependent variable y_i can in general take on two values denoted by 0 or 1. Generally it is assumed that the variables take on the value 1 if an event of interest occured.
- For instance if the response variable reports whether a unit decided or not to buy a new car, we could put for the i–th unit

$$
\begin{aligned}
y_i = 0 &\quad \text{if the car have not been purchased} \\
y_i = 1 &\quad \text{if the car have been purchased}
\end{aligned}
$$

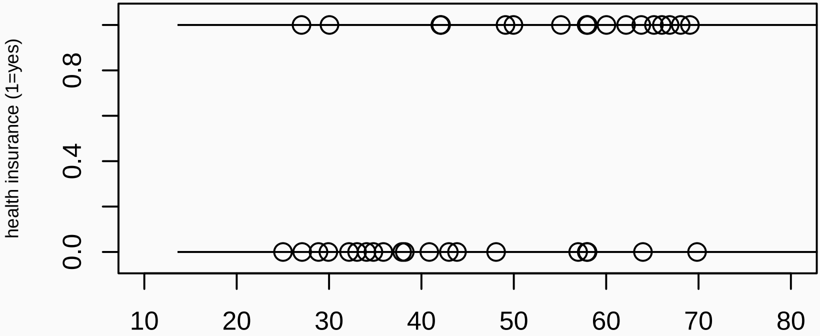
Bernoulli variables

- Variables like the one introduced above are characterized by a Bernoulli probability distribution

Y	$Pr(Y = y)$
0	$1 - p$
1	p

- p is a probability and then varies between 0 and 1.
- We expect that the probability p that a given event occurs varies according to the values of some covariates x_i.
- p is also the mean of the variable Y and so we are trying to understand if (and possibly how) the mean of the response variable varies as a function of a set of covariates.

A first example: Health Insurance coverage



For a sample of 37 individuals we observe the age of any sample unit and whether he/she owns a private health insurance.

It seems that older units are more likely to own a health insurance. For these data response variable Y can be assumed Bernoulli

1. $Y_i \sim \text{Bernoulli}(h(x_i))$.
2. and a possibly non linear model can be specified for $h(\cdot) \to [0, 1]$.

Logistic regression: Choosing an appropriate curve

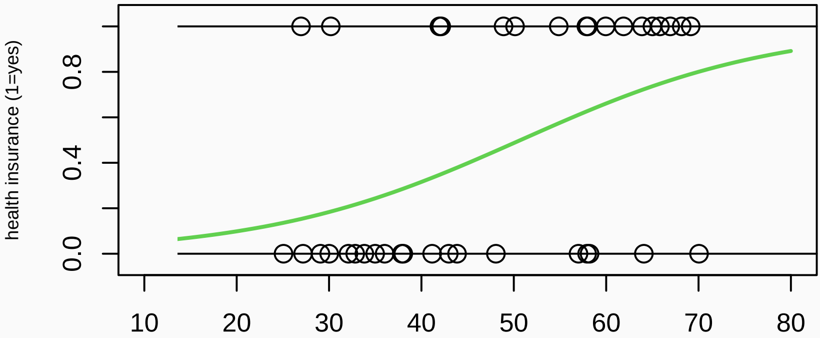
- Just like in the case of simple linear regression, our model aims to represent the mean μ_i of the dependent variable Y_i as a function of a covariate x_i
- In this case since the Y_is are drawn from a Bernoulli (or more generally Binomial) random variables, its mean is a probability.
- As we have seen an appropriate curve is not a straight line (in fact, curves that are S shaped seem more appropriate).
- There are many curves (functions) that could be considered. A possible function is the following

$$r(z) = \frac{e^z}{1 + e^z}$$

- This function, called the response function, is monotone increasing in z, exhibits an S shaped behaviour and takes on values in the interval $[0, 1]$
- Moreover, if we have a single covariate x_i we can assume that this covariate enters the function linearly, i.e.,

$$p(x_i) = r(\beta_0 + \beta_1 x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

A first example: Health Insurance coverage



The green line is the curve

$$p(eta) = g(-3.653 + 0.072eta) = \frac{e^{-3.653+0.072eta}}{1 + e^{-3.653+0.072eta}}$$

Logistic regression: Finding a “good” function

- The model defined above is the logistic regression model.
- We want to find the the parameters β_0 and β_1 that define a curve that give a better description of the data.
- Note that in this case criteria like minimization of the sum of least squares do not provide simple solutions given the non linear nature of the function r.
- But we have assumptions about which probability distribution has generated the data, more precisely we assume that:
    - for a given value of x_i we observe $y_i = 1$ with probability $p(x_i)$
    - $p(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$
    - data are independent (i.e., derived from a simple random sample of n units from a population)

Logistic regression: Maximum likelihood estimation

- Under the assumptions stated above, once data (y_i, x_i) are observed, we can evaluate what is the probability $L(\beta_0, \beta_1)\backslash$ that the observed data are generated for each possible pair of values β_0, β_1.
- The probability $L(\beta_0, \beta_1)$ is called the likelihood function and takes on different values for any possible couple (β_0, β_1).
- We could then choose that couple $\hat{\beta}_0, \hat{\beta}_1$ which corresponds to the maximum probability (maximum likelihood estimation). This couple is the maximum likelihood estimate.
- Finding the maximum likelihood estimates $\hat{\beta}_0, \hat{\beta}_1$ usually requires the use of an iterative alghorithm.

The solution obtained have "good" statistical properties especially if the sample is large

Multiple logistic regression: Extending the model

- The previous example has shown how the model can be easily entended to include more explanatory variables (in fact, we added gender).
- We can simply extend to the case where the log-odds depend linearly from a set of explanatory variables.
- This is similar to the multiple linear regression model. Then for the i–th unit in the sample we can write

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

- The covariates can be quantitative variables or indicator variables that account for qualitative factors. Also Interactions can be considered.

Structure of the model

Note that the model has a structure which is similar to the linear model

1. We specify a distributional assumption fro the responseY_i: a Bernolli variable in this case. Then $E(Y_i) = p_i$
2. We specify the way the inputs (the covariates) are combined in order to measure their impact on the expected value p_i: it is a linear combination

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

3. We specify how the linear combination η_i is related to p_i. In the case of logistic regression $r(\eta_i) = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_j x_{ij}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_j x_{ij}}} = p_i$ so that the inverse function, called the **link function** is also defined $g(p_i) = log(\frac{p_i}{1-p_i}) = \boldsymbol{x}_i^T \beta$

Maximum likelihood in details

We have a precise idea of the distribution of the response variable and we will also assume that a random sample of size n is available.

The log-likelihood $log(L(\beta)) = l(\beta)$ is

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{n} [y_i log(\pi_i) - y_i log(1-\pi_i) + log(1-\pi_i)] \\ &= \sum_{i=1}^{n} \left[y_i log(\frac{\pi_i}{1-\pi_i}) + log(1-\pi_i) \right]\end{aligned}$$

Logistic regression implies $log(\frac{\pi_i}{1-\pi_i}) = \boldsymbol{x}_i^T \beta$ and then

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \boldsymbol{x}_i^T \beta - log(1 + exp(\boldsymbol{x}_i^T \beta))] = \sum_{i=1}^{n} [y_i \eta_i - log(1 + exp(\eta_i))]$$

Equating to 0 the first derivative of $\ell(\beta)$ we obtain the likelihhod equations

$$s(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \pi_i) = 0$$

It is a system of of p non linear equations whose solution requires numerical methods.

Parameters interpretation

Logistic regression

- The interpretation of the parameters of a logistic regression model is slightly different compared with linear regression. Let us consider a simple regression model with just one variable
- The intercept, β_0, is meaningful only if $x = 0$ makes sense in the context considered.
- In the simple model here introduced, the important parameter is the one associated with the *j*-th covariate: β_j
    - if β_j is positive the larger is x the higher will be the probability that the event occurs
    - if β_j is negative for large values of x the probability that the event occurs will be lower
    - $\beta_j = 0$ implies no effect of X on the probability of the event

Logistic regression

- The parameter $\beta - J$ of a logistic regression, unlike linear regression, cannot be interpreted as the variation in the probability corresponding a variation of 1 unit in X_j
- In fact, considering for simplicity the case with a single input X, the slope of the curve is different for different values of X (since the relationship is not linear)
- but
  - we can consider the inverse of relationship $p(x_i) = g(\beta_0 + \beta_1 x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$
  - in this case we obtain $log \frac{p(x_1)}{1-p(x_1)} = \beta_0 + \beta_1 x_i$
  - $\log \frac{p}{1-p}$ is the so called logit transform of a probability p
- In the logistic regression model (or logit model) we assume that X affects linearly the logit $log \frac{p}{1-p}$

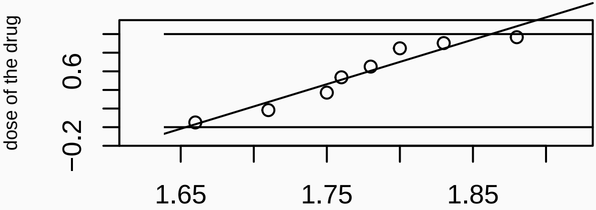
A second example: A dose-response analysis

- Consider the data in the table below

dose	1.66	1.74	1.75	1.76	1.78	1.80	1.86	1.88
n. positive	3	9	23	30	46	54	59	58
n. of patients	59	60	62	56	63	59	62	60
proportion	0.051	0.150	0.371	0.536	0.730	0.915	0.951	0.967

- The data refer to 481 individuals who received a drug. For each dose of the drug it has been observed if the individual had a positive response or not.
- Since only 8 different doses have been considered we can obtain the proportion positive responses for each dose.

Binomial response



- The plot shows that the proportion of positive responses out of m_i on trial, increases with the dose of the drug.
- A linear relationship is patently inappropriate. The data are proportions and their values should lie in the [0,1] range
- $Y_i \sim \text{Binomial}(m_i, h(x_i))$. Specify a non linear model for $h(\cdot) \to [0, 1]$.

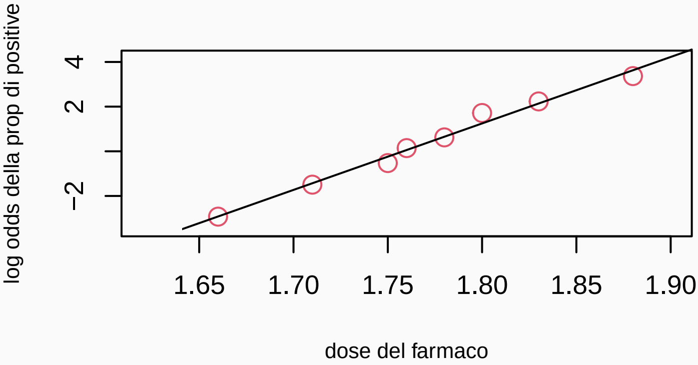
Logistic regression: The logit transform

Let us consider again the data about the proportion of positive responses to the drug.

dose	1.66	1.74	1.75	1.76	1.78	1.80	1.86	1.88
n. positive	3	9	23	30	46	54	59	58
n. of patients	59	60	62	56	63	59	62	60
proportion (p)	0.051	0.150	0.371	0.536	0.730	0.915	0.951	0.967
$p/(1-p)$	0.05	.177	0.59	1.15	2.71	10.80	19.67	29.00
$log(p/(1-p))$	-2.92	-1.73	-0.53	0.14	0.99	2.38	2.98	3.36

- $\frac{p}{1-p}$ are the odds. Odds provide an alternative way to descrive the probability of an event. They take on values between 0 and ∞

Logistic regression: Alternative representation of the dose response model



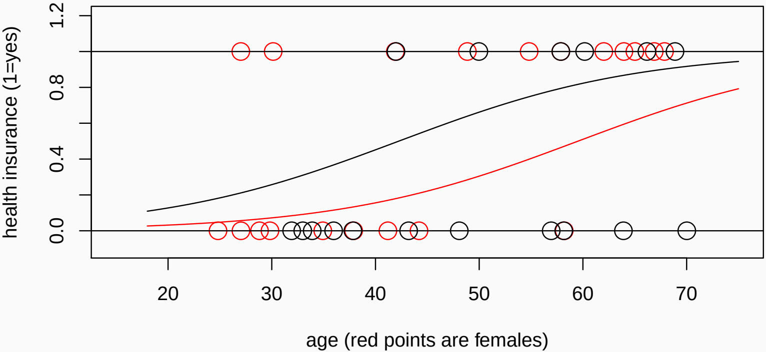
- The relationship between dose and log–odds of the proportion is linear!
- This means that a unit increase of the dose will cause an increase of β_1 in the log–odds of the proportions

Logistic regression: Odds and log–odds

- Bernoulli random variables are completely defined by the value of p, the probability of a "success". The odds defined as $\frac{p}{1-p}$, obtained by a simple transfomation of p, have an important interpretation.
- Suppose p indicates whether a given football team wins the next match. If $p = 0.2$ than the odds of the team winning are 0.2/(1-0.2)=1/4 and we may say that the odds of winning are 1 on 4.
- This means that if we bet 1 euro on the team winning, in a fair game, if the team wins we get the euro back plus 4 euros. If the team does not win, we lose our euro.
- The odds provides the important information in this context (bet of 1 and winning of 4) and in fact when betting the information provided are simply the odds.
- If we know the odds we can calculate the probability p and vice versa.
- The odds can take on any positive value and the odds are 1 when an event has probability $p = 0.5$.
- The logarithm of the odds is often used, it can take any value and it is equal to 0 if the probability $p = 1/2$.
- As we have noted β_1 in our simple logistic regression model is the proportional variation we observe in the log-odds if the covariate X is increased by a unit.

Interpretation of a dichotomous covariate: Health Insurance coverage continued

- Let us consider again the data on private health insurance and assume we know observe the gender of the respondents



Logistic regression with a dichotomous covariate: Health Insurance coverage continued

- This is the result for a more complex logistic regression model

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x + \beta_2 sex$$

  sex can take on only two values 0 (if female) or 1 (if male)

- The maximum likelihood estimates of the coefficients are

(Intercept)	eta	sex
-5.152	0.087	1.496

- Probability of owing a health insurance is higher for males and increases with age

Logistic regression with a dichotomous covariate: Health Insurance coverage continued

- If we evaluate the difference in the log-odds of the probability of health insurance (at a given age) for males, p_{male}, and females, p_{female}, this will be simply equal to 1.496
- $log\frac{p_{male}}{1-p_{male}} - log\frac{p_{female}}{1-p_{female}} = 1.496$
- or equivalently $log\frac{\frac{p_{male}}{1-p_{male}}}{\frac{p_{female}}{1-p_{female}}} = 1.496$
- The estimated coefficient $\beta_2 = 1.496$ represents the so called log–odds ratio
- And $e^{1.496}$ is the odds ratio
- Odds ratio is 1 if the the two odds (or) the two probabilities are the same for males and female

Logistic regression with a dichotomous covariate: Health Insurance coverage continued

- Log-odds ratio is 0 if the two probabilities are the same ...
- and when the probability of a health insurance is the same for males and females then having or not a health insurance policy do not depend on the gender.
- In this case the value of $\beta_2 = 1.496$ indicates a seemingly not negligible change in the log-odd ratio and it means that probability is different for males and female.
- The odds ratio $e^{\beta_2} = e^{1.496} = 4.464$ indicates that the odds of having a health insurance for a male are more than 4 times the same odds for a female.

Males are about 4.5 times more likely to have a health insurance policy than females.

Inference for logistic regression parameters

Testing parameters significance

- Maximum likelihood method provides good estimates of the βs.
- For the j–th variable X_j we want to state if the data convey enough evidence to draw the conclusion that this variable is relevant to predict the response variable.
- Maximum likelihood methods provides also estimates of the standard errors of the estimated parameters.
- For (moderately) large sample we are able to answer to the question:

"is a given parameter significantly different from zero?"

or stated more formally, we want to test the hypothesis

$$H_0 : \beta_j = 0$$

Testing parameters significance

- As in the linear regression case we can consider the ratio
$$z = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$$
- This ratio, if the hypothesis $H_= : \beta_j = 0$ holds, should be a value from a $N(0, 1)$. If absolute value of z is too "large" to believe it is a value from a standard Normal distribution, then data do not support the hypotesis that the parameter is zero;
- then to decide when "large" is really large, one can give a look to the associated p–values. This is the probability that we obtain a z even larger that the one observed when the parameter is actually equal to 0.
- since p–values are probabilities, they lies between 0 and 1. And usually one judge the j–th variable relevant if the p–value associated to its estimate is (possibly much) smaller than 0.05.

Testing parameters significance

1. The result above follows from the asymptotic properties of MLE: for large n we know that $\hat{\beta} \dot{\sim} \mathcal{N}(\beta, I(\beta)^{-1})$ where $I(\beta)$ is the expected information matrix, which in the case of a Bernoulli model is

$$I(\beta) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \pi_i (1 - \pi_i)$$

   where $\pi_i = r(\boldsymbol{x}_i^T \beta)$.

2. This matrix depends on the unknown quantities β but a consistent estimates is obtained but substituting to β its estimate $\hat{\beta}$.
3. The element on the diagonal of $I(\beta)_{jj}^{-1}$ is an estimate of the variance of $\hat{\beta}_j$.
4. For this reason the ratio $\frac{\beta_j}{\sqrt{I(\hat{\beta})_{jj}^{-1}}}$ evaluated , is asymptotically distributed as a Standard Gaussian assuming $H_0 : \beta_j = 0$.

Inference for logistic regression parameters: Judging the overall performance of the model

- For the logistic regression model it is not possible to obtain a quantity that has the same interpretation of R^2 in the linear model.
- It is possible to measure the difference between the value of the likelihood for the estimated parameters $L_{\hat{\beta}} = L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_j)$ and the value of the likelihood we would obtain in other cases.
- Two relevant cases are
  - the likelihood L_{max} one could achieve if considers as many parameters as available data (thus achieving a perfect fit)
  - the likelihood L_0 one obtains in a null model , i.e., a model with only the intercept β_0 (this means that no covariate has a sugnificant effect on the response).
- Comparing those likelihoods helps to judge whether the model is useful to predict the response variable

Inference for logistic regression parameters: Judging the overall performance of the model

- It is possible to look at the ratio between $L_{\hat{\beta}}$ and L_0 or at the difference between $logL_{\hat{\beta}}$ and $logL_0$:
  if the latter difference is small then the model is not supported by the data
- It is also possible to consider the difference between the $logL_{max}$ and $logL_{\hat{\beta}}$. This difference should be small for good models.
- The value $D = 2(logL_{max} - logL_{\hat{\beta}})$ is called the deviance.
- It behaves like the deviance in the linear model: is large for bad models and decreases as we improve the model for instance by adding more significant explanatory variables.
- Comparing the deviances of two alternative models that differ only because a simpler model is obtained by setting some parameters equal to 0 (i.e. excluding some potential covariates) helps to decide which one among the two models should be preferred.

Logistic regression results: Health Insurance coverage

```r
mod1<-glm(formula = sani ~ eta + sex, family = binomial(link=logit))
summary(mod1)
```

```
##
## Call:
## glm(formula = sani ~ eta + sex, family = binomial(link = logit))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.15175    1.79715  -2.867  0.00415 **
## eta          0.08654    0.03128   2.767  0.00567 **
## sexm         1.49569    0.85484   1.750  0.08017 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 51.049  on 36  degrees of freedom
## Residual deviance: 39.612  on 34  degrees of freedom
## AIC: 45.612
##
## Number of Fisher Scoring iterations: 4
```

Logistic regression: Predicting the response variable

- Remind that in a logistic regression model we assume that

$$p_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_j x_{ij}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_j x_{ij}}}$$

- We can simply estimate the probabilities p_i by substituting the estimated values to the βs

$$\hat{p}_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}+\cdots+\hat{\beta}_j x_{ij}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}+\cdots+\hat{\beta}_j x_{ij}}}$$

- These predicted probabilities are used when this model is used for classification. Simply define a threshold $c \in (0, 1)$ and predict $Y_i = 1 \quad \text{if} \quad \hat{p}_i > c$

Alternative specification of the response function

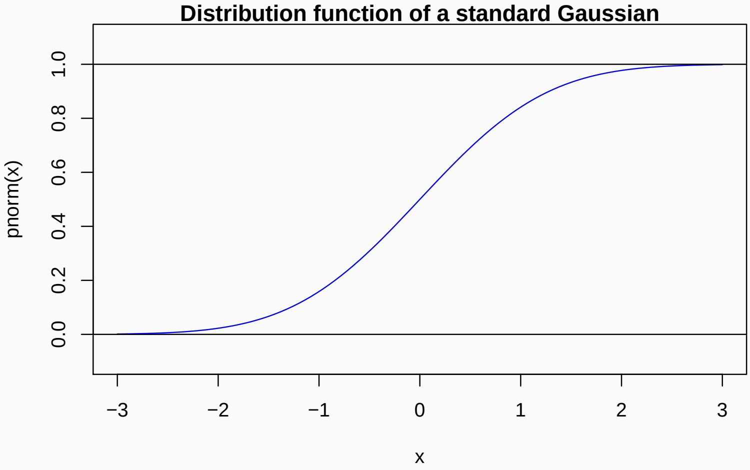
Probit regression

- We justified the choice of the response function $g(z)$ that gave rise to logistic regression by saying that we needed a S shaped function that lies within the $[0, 1]$ range since we want it to represent probabilities.
- But there are many function that we could choose. For instance a function that could work well is the distribution function of the standard Gaussian
- In fact we could write

$$p_i = \Phi(\beta_0 + \beta_1 x_i)$$

where the function Φ is the distribution function of the standard Gaussian random variable

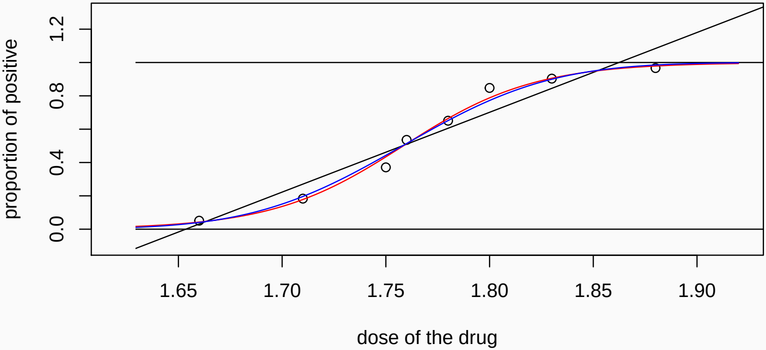
Probit regression



- This choice of the response function defines the probit regression model
- Probit regression model is also very popular
- Other choices are also possible for $g(.)$

Probit vs logistic regression

- Actually probit regression gives results that are very similar to those obtained with logistic regression



the blue curve represents prediction by a probit regression model

Estimation issues

The case of perfect separation

- The maximum likelihood estimates for a binomial model are generally easily found using efficient numerical algorithms
- However, there may be convergence problems if it exist a function of the covariates that perfectly separates $y_i = 1$ and $y_i = 0$. Or if for some categories defined by a covariate, y to is only 0 or only 1.
- In this case the likelihood function does not have a maximum and as a results the estimates provided are highly unstable.
- The main symptom is therefore given by a message that says "the algorithm has not reached convergence" and that "probability predictions have been obtained which are numerically equal to 1 or 0". Another symptom is that the values of the standard errors of the estimates are very high.
- There are several solutions. One possible solution is the one that uses a penalized likelihood.
- This solution can be obtained by considering a likelihood to which a term is added to eliminate the bias in ML estimates for logistic regression (for example by using the `{R brglm}` package)