UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

Statistics

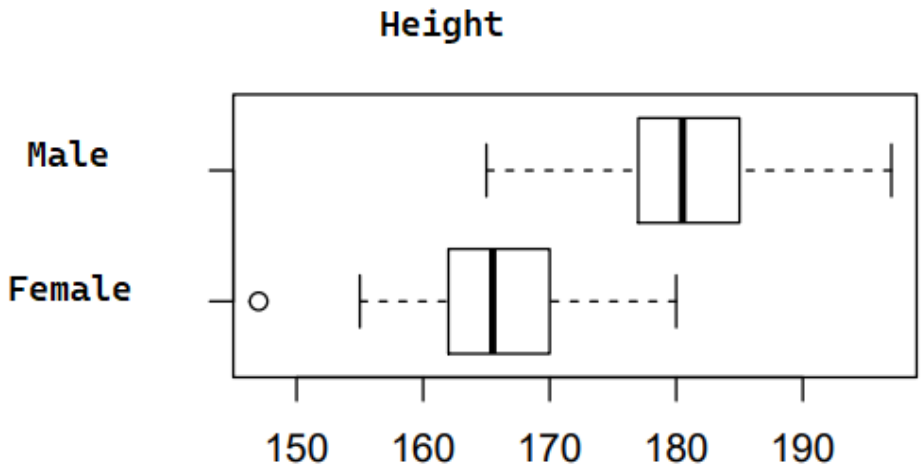
Logistic Regression

Sara Geremia

May 9th, 2024

Let's start from here:

Height and gender (in a set of units) are associated



In different application contexts, there is interest in studying the relationship between a dichotomous dependent variable Y and one or more independent variables

Binary Response Variable

The analysis of a binary response variable has required the development of a methodology that is generalizable to this type of data and comparable to the use of linear models for normally distributed variables.

$$y = \beta_0 + \beta_1 x + \text{error}$$

The success of linear analyses, and thus of the least squares estimation method, stems from the reasonableness and simplicity of a linear model and the properties enjoyed by the least squares estimators under appropriate assumptions.

Binary Response Variable

If the dependent variable is a binary variable, the expected value of the variable Y has a directly interpretable meaning: it is the conditional probability of a positive outcome.

$$\mathbb{E}(Y|X) = P(Y = 1|X)$$

The assumptions introduced for the linear regression model are not satisfied. The variable Y and the error term are neither normally distributed nor have constant variance.

Moreover, it does not restrict $P(Y = 1|X)$ to lie between 0 and 1.

Nonlinear model

We need an approach that uses a nonlinear function to capture the relationship between the predictors and the probability of the event.

Commonly used methods for this purpose are Probit and Logit regression.

The probit and logit transformations are used to model the probability of the event $P(Y = 1|X)$ in a way that is suitable for binary outcomes.

Probit Regression

In Probit regression, the cumulative standard normal distribution function $\Phi(\cdot)$ is used to model the regression function:

$$P(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$$

$\beta_0 + \beta_1 X$ plays the role of a quantile z.

such that the coefficient β_1 is the change in z associated with a one-unit change in X. While the link between z and Y is non linear, since Φ is a non-linear function.

Logit Regression

In Logit regression, the standard cumulative logistic distribution function $F(x) = \frac{1}{1+e^{-x}}$ is used to model the regression function:

$$P(Y = 1|X) = F(\beta_0 + \beta_1 X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}} = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

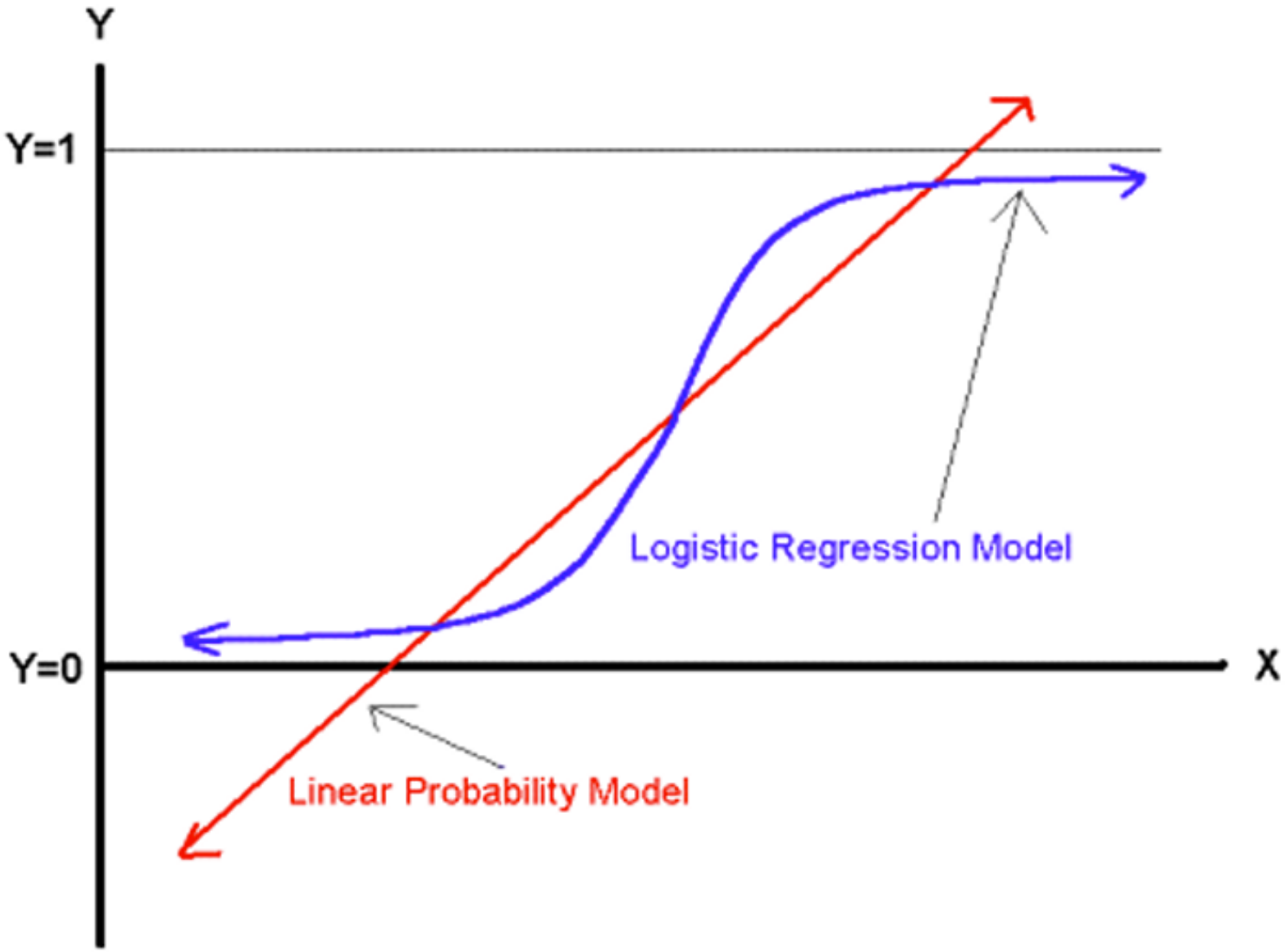
Logit Regression

$F(\cdot)$ has characteristics similar to the standardized normal.

The curve described by the logistic model has the following properties:

- β_1 coefficient measure how fast $P(Y|X)$ varies from 0 to 1.
- As X increases, the values tend towards one if $\beta_1 > 0$.
- It is a monotonic function, meaning the curve is increasing (or decreasing) everywhere.

Logit Regression



Logit Regression

According to the model described, the probability of a positive response (Y = 1) as a function of the the variable X is:

$$P(Y|X) = \frac{e^{\beta_0+\beta X}}{1+e^{\beta_0+\beta X}}$$

while the probability of a negative response is:

$$1 - P(Y|X) = \frac{1}{1+e^{\beta_0+\beta X}}$$

The odds of a positive response, conditioned on X, is given by:

$$\text{Odds}(Y|X) = \frac{P(Y|X)}{1-P(Y|X)} = \frac{\frac{e^{\beta_0+\beta X}}{1+e^{\beta_0+\beta X}}}{\frac{1}{1+e^{\beta_0+\beta X}}} = e^{\beta_0+\beta X}$$

Odds

A quantity commonly used to study binary variables is the Odds.

$$\text{Odds}(Y|X) = \frac{P(Y|X)}{1 - P(Y|X)} \in [0, \text{inf}]$$

It represent the ratio of the probability of an event occurring to the probability of it not occurring. The odds reflect how much success is more likely than failure. If the odds are 1, it implies that the event is equally likely to occur as not to occur.

For instance, if the odds of success are 2, it means that the probability of success is twice as likely as the probability of failure.

Logit Transformation

The logarithm of the odds is therefore:

$$\ln\left(\frac{P(Y|X)}{1-P(Y|X)}\right) = \beta_0 + (\beta_1 X)$$

The logit transformation is defined as:

$$\text{logit}(P(Y|X)) = \ln\left(\frac{P(Y|X)}{1-P(Y|X)}\right)$$

The logit transformation allows us to reformulate the previously described model in terms of a linear regression model. The logit transformation linearizes the model but does not eliminate the problem of heteroscedasticity and other issues, making it appropriate to resort to a different estimation method, namely the maximum likelihood estimation method.

Odds Ratio (OR)

A very important characteristic of the logistic regression model is related to the interpretation of the coefficients.

For a continuous independent variable the odds ratio can be defined as:

$$OR = \frac{odds(Y|X = x + 1)}{odds(Y|X = x)} = \frac{e^{\beta_0+\beta_1(X+1)}}{e^{\beta_0+\beta_1 X}} = e^{\beta_1}$$

e^{β_1} is the odds ratio associated with a one-unit increase in X.

Odds Ratio (OR)

The odds ratio is generally the parameter of greatest interest for interpreting the coefficients.

Particularly in epidemiological studies, depending on the value of β_1, there will be a different interpretation of the coefficients:

- $\beta_1 = 0$: There is independence between Y and X; the exposure is not associated with the outcome (OR=1).
- $\beta_1 < 0$: X is a protective factor for a certain disease (represented by Y) (OR¡1).
- $\beta_1 > 0$: X is a risk factor for Y (OR¿1).

Odds Ratio (OR)

Sometimes it's not interesting to consider a unitary increase, but it's preferable to consider a variation Δ of the independent variable X.

The corresponding odds ratio is:

$$\text{Ex. } \text{logit}\left(\pi(Y|X)\right) = \beta_0 + \beta_1 X\Delta$$

For example, in a study examining gender (M/F) as the outcome and height as the explanatory variable, the odds ratio $\exp(5 \cdot \beta_1)$ would be derived from comparing two individuals differing in height by 5 units (e.g., 5 centimeters). It represents the odds ratio for gender comparing a group of subjects who are 5 units taller with a group who are 5 units shorter.