UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

Statistics

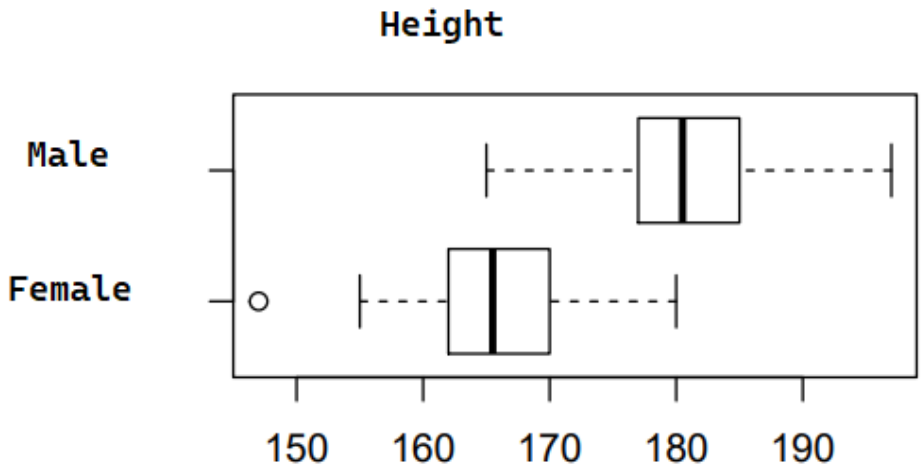
Logistic Regression

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:

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

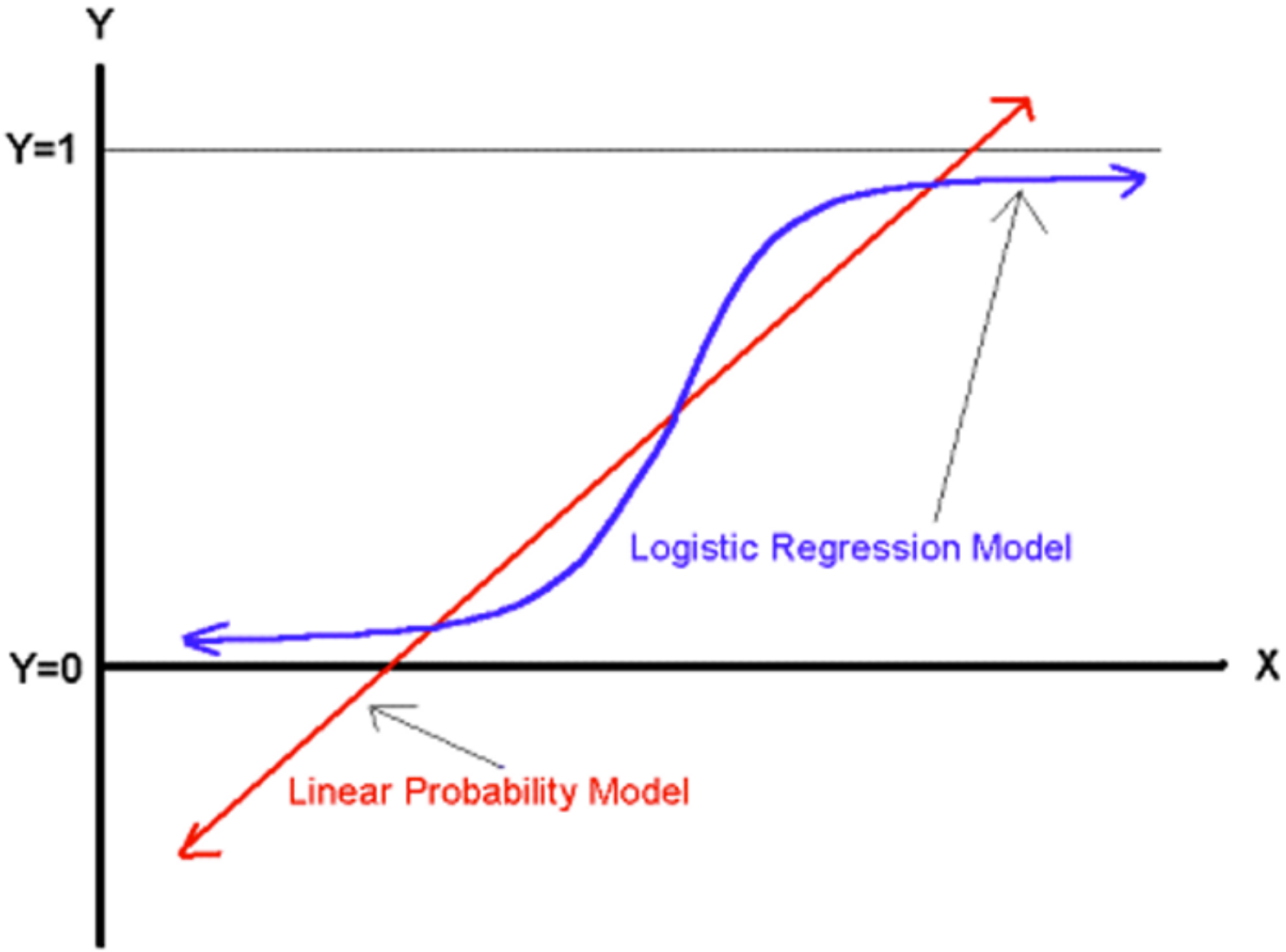
Logit Regression

$F(\cdot)$ 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:

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

while the probability of a negative response is:

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

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

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

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): Continuous X

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): Continuous X

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): Continuous X

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:

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

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.

Odds Ratio (OR): Binary X

	X = 1	X = 0
Y = 1	$P(Y\|X=1)$	$P(Y\|X=0)$
Y = 0	$1-P(Y\|X=1)$	$1-P(Y\|X=0)$

The odds is:

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

Hence,

$$Odds(Y|X=1) = \exp^{\beta_0+\beta_1}, \qquad Odds(Y|X=0) = \exp^{\beta_0}$$

Odds Ratio (OR): Binary X

	X = 1	X = 0
Y = 1	$P(Y\|X=1)$	$P(Y\|X=0)$
Y = 0	$1-P(Y\|X=1)$	$1-P(Y\|X=0)$

$$Odds(Y|X=1) = \exp^{\beta_0+\beta_1}, \qquad Odds(Y|X=0) = \exp^{\beta_0}$$

$$OR = \frac{Odds(Y|X=1)}{Odds(Y|X=0)} = \frac{\exp^{\beta_0+\beta_1}}{\exp^{\beta_0}} = \exp^{\beta_1}$$

$$ln(OR) = \beta_1$$

Odds Ratio (OR): Categorical X with K levels

For example, age considered in three intervals

$$
\begin{aligned}
&< 50\ \textit{years},\\
50 - &59\ \textit{years},\\
&60 +\ \textit{years}
\end{aligned}
$$

We need to use a set of variables that allows us to represent the three categories. Let's consider two (k-1) dummy (binary) variables, D1 and D2.

	D1	D2
$< 50\ \textit{years}$	0	0
$50 - 59\ \textit{years}$	1	0
$60 +\ \textit{years}$	0	1

Odds Ratio (OR): Categorical X with K levels

	D1	D2
< 50 *years*	0	0
$50 - 59$ *years*	1	0
$60 +$ *years*	0	1

$$\text{Odds}(Y|X =< 50) = e^{\beta_0 + \beta_1(D1=0) + \beta_2(D2=0)}$$

$$\text{Odds}(Y|X = 50 - 59) = e^{\beta_0 + \beta_1(D1=1) + \beta_2(D2=0)}$$

$$\text{Odds}(Y|X = 60+) = e^{\beta_0 + \beta_1(D1=0) + \beta_2(D2=1)}$$

Odds Ratio (OR): Categorical X with K levels

	D1	D2
< 50 *years*	0	0
$50 - 59$ *years*	1	0
$60 +$ *years*	0	1

$$\text{OR} = \frac{\text{Odds}(Y|X = 50-59)}{\text{Odds}(Y|X =< 50)} = \frac{e^{\beta_0+\beta_1(D1=1)+\beta_2(D2=0)}}{e^{\beta_0+\beta_1(D1=0)+\beta_2(D2=0)}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}$$

$$\text{OR} = \frac{\text{Odds}(Y|X = 60+)}{\text{Odds}(Y|X =< 50)} = \frac{e^{\beta_0+\beta_1(D1=0)+\beta_2(D2=1)}}{e^{\beta_0+\beta_1(D1=0)+\beta_2(D2=0)}} = \frac{e^{\beta_0+\beta_2}}{e^{\beta_0}}$$

The age group < 50 is the reference category for both coefficients.

Odds Ratio (OR): Categorical X with K levels

If we want to compare the third and the second category, we need to consider the difference between the corresponding regression coefficients.

$$\text{OR} = \frac{\text{Odds}(Y|X = 60+)}{\text{Odds}(Y|X = 50-59)} = \frac{e^{\beta_0+\beta_1(D1=0)+\beta_2(D2=1)}}{e^{\beta_0+\beta_1(D1=1)+\beta_2(D2=0)}} = e^{\beta_2-\beta_1}$$

Odds Ratio (OR): Categorical X with K levels

Let's say the dependent variable Y represents the likelihood of developing a certain disease.

- **Reference Category Selection**: Let's choose $< 50 years$ as the reference category.
- **Interpretation of ORs**:
  - *50-60 years*: An OR greater than 1 suggests that individuals in the 50-60 age group are more likely to develop the disease compared to those under 50 years old. Conversely, an OR less than 1 suggests they are less likely to develop the disease compared to the reference group.
  - *60+ years*: An OR greater than 1 suggests that individuals aged 60 or above are more likely to develop the disease compared to those under 50 years old.
- **Magnitude of Association**: Higher OR values indicate a stronger association.
- **Statistical Significance**: A p-value less than your chosen significance level (e.g., 0.05) indicates that the association is statistically significant.

Multivariate logistic regression

Multivariate logistic regression model:

$$\text{logit}\left(P(X_1, X_2)\right) = \beta_0 + \beta_1 X_1 \beta_2 X_2$$

Where:

- Y = Coronary Heart Disease (present/absent)
- X_1 = Age; X_2 = Smoking

$\exp(\beta_1)$ is the Odds Ratio (OR) comparing two individuals differing by one year in age, both being either smokers or non-smokers. It's an OR for CHD associated with the age variable, adjusted for smoking.

$\exp(\beta_2)$ is the Odds Ratio comparing two individuals of the same age, one being a smoker and the other a non-smoker. It's an OR for CHD associated with the smoking variable, adjusted for age.

Multivariate logistic regression

The logit function

$$\text{logit}(P(Y|X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

estimates the relationship intensity between each predictor and the outcome after adjusting for all other variables in the model. The logistic regression model is a multiplicative model of risks.

We introduce the **multivariate logistic regression model** to better understand the relationship between predictors and the outcome, considering the potential influence of omitted factors and providing more robust estimates of their effects on the outcome variable.

Simple vs Multivariate Regression

- One dependent variable Y predicted from one independent variable X
- One regression coefficient
- R^2: proportion of variation in dependent variable Y predictable from X

- One dependent variable Y predicted from a set of independent variables $(X_1, X_2 \ldots . X_p)$
- One regression coefficient for each independent variable
- R^2: proportion of variation in dependent variable Y predictable by set of independent variables (X's)

Significance Tests

Testing R2

- Test R^2 through an F test
- Test of competing models (difference between R^2) through an F test of difference of R^2s

Testing β

- Test of each partial regression coefficient (β) by t-tests

Different Ways of Building Regression Models

- Simultaneous: All independent variables entered together
- Stepwise: Independent variables are removed or included ain order to find the subset of variables in the data set resulting in the best performing model
- Hierarchical: variables are entered into the model in predefined stages or blocks. Each stage represents a group of variables that are theoretically related or logically connected.

Stepwise Regression

The stepwise regression consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error.

- **Forward selection**: starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.
- **Backward selection**: starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.
- **Stepwise selection**: is a combination of forward and backward selections. You start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection).