Data Valorization:
Logistic Regression

Lionel Fillatre

fillatre@unice.fr

Outline

- Introduction
- Odds and logit
- Interpretation
- Maximum Likelihood
- Conclusion

1 Introduction

General Linear Models

- Family of regression models

- **Response** / **Model Type**

Response	Model Type
Continuous	Linear regression
Counts	Poisson regression
Survival times	Cox model
Binomial	Logistic regression

- Uses
  - Model building, risk prediction, etc.

Logistic Regression

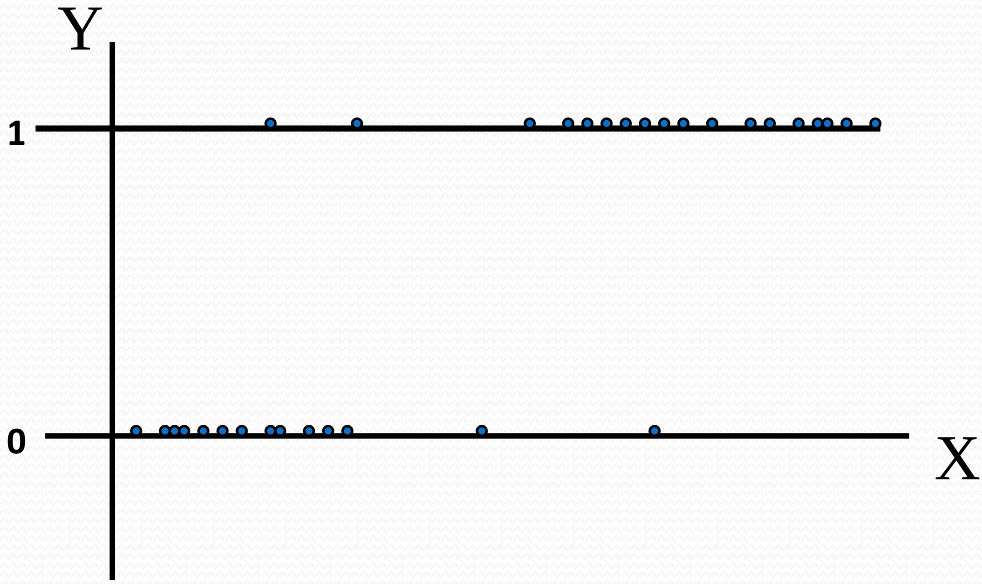
- Models relationship between set of variables X_i
  - dichotomous (yes/no, smoker/nonsmoker,…)
  - categorical (social class, race, ...)
  - continuous (age, weight, gestational age, ...)

- And a dichotomous categorical response variable Y

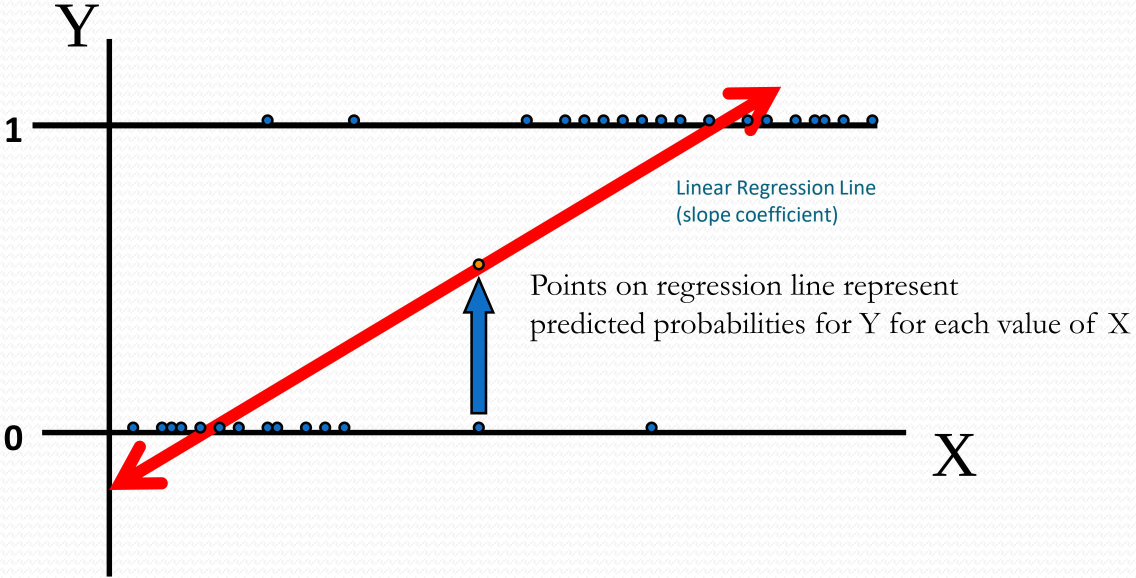
  e.g. Success/Failure, Remission/No Remission, Survived/Died, etc…

Scatterplot with Y=(0,1)

Y = Hired-Not Hired; X= Experience



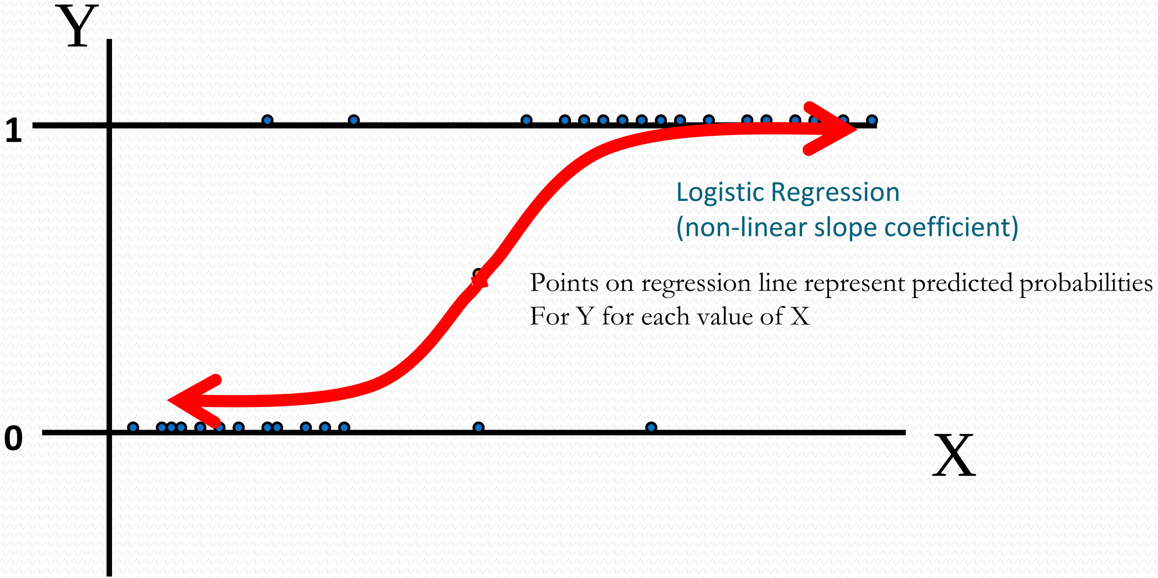
Simple Linear Regression



Binary Logistic Regression or “Logit”

- Selects regression coefficient to force predicted values for Y to be between (0,1)
- Produces S-shaped regression predictions rather than straight line
- Selects these coefficient through “Maximum Likelihood” estimation technique

Picture of Logistic Regression



Requirements for Logistic Regression

The Following need to be specified:

1) An outcome variable with two possible categorical outcomes (1=success; 0=failure).

2) A way to estimate the probability P of the outcome variable success.

3) A way of linking the outcome variable to the explanatory variables.

4) A way of estimating the coefficients of the regression equation.

2 Odds and Logit

Definition of odds

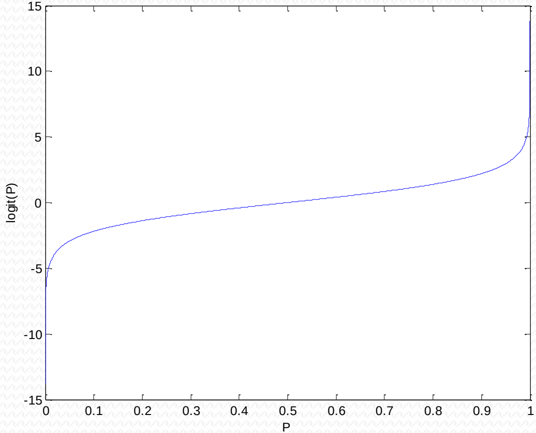
- Let P the probability that Y takes the value 1
- Since $0 \leq P \leq 1$, we might use odds $= \frac{P}{1-P}$
- The natural way to interpret odds for is as the ratio of events to non-events in the long run.
  - Example: odds for rolling six with a fair die are 1 to 5. This is because, if one rolls the die many times, and keeps a tally of the results, one expects 1 six event for every 5 times the die does not show six.
- Odds has no “ceiling” but has “floor” of zero.

Definition of logit

- So we use the logit transformation

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

- Logit does not have a floor or ceiling.

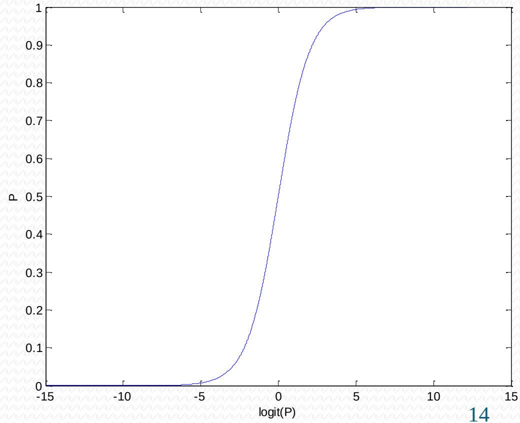


Inverse of logit: Logistic Function

- Since odds $= P/(1-P)$, it follows that $P = \text{odds}/(1 + \text{odds})$.
- Besides, $\ln(\text{odds}) = \text{logit}(P)$ involves odds $= e^{\text{logit}(P)}$, hence

$$P = \frac{e^{\text{logit}(P)}}{1 + e^{\text{logit}(P)}} = \frac{1}{1 + e^{-\text{logit}(P)}} = f\big(\text{logit}(P)\big)$$

where $f(z) = \frac{1}{1+e^{-z}}$ is called the logistic function.



Logistic Regression Model

- Assumption: $P = P(Y = 1|X) = P(X) = P(X_1, X_2, \ldots, X_k)$

- Logistic Regression Model (LOGIT Transform):

$$\text{logit}\big(P(X)\big) = \ln\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

or

$$\text{odds}(X) = e^{\text{logit}(P(X))} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}$$

- The coefficients β_0, β_1,..., β_k should be estimated (discussed later).

3 Interpretation

Interpretation

- $\text{logit}(P(X)) \;=\; \ln\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$
- β_0 represents the global factor
- β_1 represents the fraction by which the risk is altered by a unit change in X_1
- β_2 is the fraction by which the risk is altered by a unit change in X_2
- And so on.
- What changes is the log odds; the odds themselves are changed with an exponential factor.

Interpretation

- If $\ln(\text{odds}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$ then

$$\text{odds} = (e^{\beta_0})\,(e^{\beta_1 X_1})\,(e^{\beta_2 X_2})\ldots(e^{\beta_k X_k})$$

- Model is multiplicative on the odds scale

Example: Dichotomous Predictor

- Consider a dichotomous predictor (X) which represents the presence of risk (1 = present)

	Risk Factor (X)	
Disease (Y)	**Present** (X = 1)	**Absent** (X = 0)
Yes (Y = 1)	$P(Y=1 \mid X=1)$	$P(Y=1 \mid X=0)$
No (Y = 0)	$1-P(Y=1 \mid X=1)$	$1-P(Y=1 \mid X=0)$

$$\frac{P(Y=1 \mid X)}{1-P(Y=1 \mid X)} = e^{\beta_o+\beta_1 X} \quad \begin{cases} \text{Odds for Disease with Risk Present} = \dfrac{\mathrm{P}(\mathrm{Y}=1 \mid \mathrm{X}=1)}{1-\mathrm{P}(\mathrm{Y}=1 \mid \mathrm{X}=1)} = e^{\beta_o+\beta_1} \\ \text{Odds for Disease with Risk Absent} = \dfrac{\mathrm{P}(\mathrm{Y}=1 \mid \mathrm{X}=0)}{1-\mathrm{P}(\mathrm{Y}=1 \mid \mathrm{X}=0)} = e^{\beta_o} \end{cases}$$

Therefore the odds ratio (OR) $= \dfrac{\text{Odds for Disease with Risk Present}}{\text{Odds for Disease with Risk Absent}} = \dfrac{e^{\beta_o+\beta_1}}{e^{\beta_o}} = e^{\beta_1}$

Interpretation of odds ratio (binary case)

- $\mathrm{OR}_i = 1$: the « success » is independent of the variable X_i
- $\mathrm{OR}_i > 1$: the « success » occurs more often for samples when X_i is true
- $\mathrm{OR}_i < 1$: the « success » occurs more often for samples when X_i is false

Example: Odds Ratio

- P is proportion of individuals with a Myocardial Infarction
-
- Predictors:
  - age in years
  - htn = hypertension ($1 =$ yes, $0 =$ no)
  - smoke = smoking ($1 =$ yes, $0 =$ no)

- Estimated model:
  - Logit$(P) \;=\; \beta_0 + \;\beta_1$ age $+\;\beta_2$ htn $+\;\beta_3$ smoke

- Question : want OR for a 40 year old with hypertension vs otherwise identical 30 year old without hypertension.

- Answer: $\beta_0 + \beta_1\, 40 + \beta_2 + \beta_3$ smoke– $(\beta_0 + \beta_1\, 30 + \beta_3$ smoke)

  $= \beta_1\, 10 + \beta_2 = \log \mathrm{OR} \quad \Rightarrow \quad \mathrm{OR} = e^{10\beta_1 + \beta_2}$

4 Maximum Likelihood

Constructing the estimated model

- Training data set with N samples $\left(y_i, X_i = (X_{i,1}, \ldots, X_{i,k})\right)$:

$$\left\{\left(y_1, (X_{1,1}, \ldots, X_{1,k})\right), \left(y_2, (X_{2,1}, \ldots, X_{2,k})\right), \ldots, \left(y_N, (X_{N,1}, \ldots, X_{N,k})\right)\right\}$$

- The joint conditional probability of the training labels is

$$\Pr(y_1, y_2, \ldots, y_N | X_1, \ldots, X_N) = \Pr(y_1|X_1)\Pr(y_2|X_2) \cdots \Pr(y_N|X_N) = \prod_{i=1}^{N} \Pr(y_i|X_i)$$

Assumption: independent y_i's

Maximum Likelihood (ML)

- Recall that
  - $\ln\left(\frac{\Pr(Y=1|X)}{1-\Pr(Y=1|X)}\right) = \operatorname{logit}\big(\Pr(Y = 1|X)\big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$
  - $\Pr(Y = 1|X) = f(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$ where f is the logistic function
- Hence, the joint conditional probability of the labels is

$$\Pr(y_1, y_2, \dots, y_N | X_1, \dots, X_N) = \prod_{i=1}^{N} \Pr_\beta(y_i|X_i)$$

where $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)$ is the vector of parameters to be estimated.

- From the ML principle, we choose parameters $\hat{\beta}$ that satisfy

$$\hat{\beta} = \operatorname*{argmax}_{\beta \in \mathbb{R}^{k+1}} \prod_{i=1}^{N} \Pr_\beta(y_i|X_i)$$

ML calculation

- Let us denote $p = p(X) = \Pr(Y = 1|X) = \Pr_\beta(Y = 1|X)$.

- Then, for dichotomous outcome, $\Pr(Y = 0|X) = 1 - \Pr(Y = 1|X) = 1 - p$.
  It is rewritten as:
  $$\Pr(y|X) = p^y(1-p)^{1-y} \quad \text{for } y \in \{0,1\}$$

- Proof:
  - For $y = 1$, $\Pr(y|X) = \Pr(Y = 1|X) = p^1(1-p)^0 = p$
  - For $y = 0$, $\Pr(y|X) = \Pr(Y = 0|X) = p^0(1-p)^1 = 1 - p$

ML calculation

- So, given that $\Pr(y_i|X_i) = p_i^{y_i}(1-p_i)^{1-y_i}$ with $p_i = \Pr(Y_i = 1|X_i)$:

$$L = \prod_{i=1}^{N} \Pr(y_i|X_i) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i}$$

$$= \prod_{i=1}^{N} p_i^{y_i}\left(\frac{1}{1-p_i}\right)^{y_i}(1-p_i)$$

$$= \prod_{i=1}^{N} \left(\frac{p_i}{1-p_i}\right)^{y_i}(1-p_i)$$

ML calculation

- Taking the logarithm of both sides:

$$\ln L = \sum_i y_i \ln\left(\frac{p_i}{1-p_i}\right) + \sum_i \ln(1-p_i)$$

- Remember that

$$\ln\left(\frac{p_i}{1-p_i}\right) = \ln\left(\frac{P(y_i|X_i,\beta)}{1-P(y_i|X_i,\beta)}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k} = \beta X_i$$

- Vector notation: $\beta X_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}$

- Substituting in using logistic regression model:

$$\ln L = \sum_i y_i \beta X_i - \sum_i \ln\left(1 + e^{\beta X_i}\right) = J(\beta)$$

Gradient Calculation

- Cost: $J(\beta) = \sum_i y_i \beta X_i - \sum_i \ln\left(1 + e^{\beta X_i}\right)$

- Unfortunately, there is no closed form solution to maximizing $J(\beta)$ with respect to β.

- Therefore, one common approach is to use **gradient ascent**.

- Gradient with respect to β:

$$\nabla_\beta J(\beta) = \sum_i y_i X_i - \sum_i \frac{e^{\beta X_i}}{1 + e^{\beta X_i}} X_i = \sum_i X_i (y_i - \hat{p}_i)$$

Gradient Ascent

- Take a guess $\hat{\beta}_0$

- Loop over $t = 1, \ldots, M$
  - Move in the direction of the gradient

$$\hat{p}_{i,t} = \frac{1}{1 + e^{-\hat{\beta}_t X_i}} = \hat{p}_{i,t}(\hat{\beta}_t, X_i)$$

$$\hat{\beta}_{t+1} = \hat{\beta}_t + \gamma \sum_i X_i (y_i - \hat{p}_{i,t})$$

- Comment: γ is the step of the ascent (to tune carefully).

6 Conclusion

Conclusion

- Very useful for binary outcomes
- The logistic function is often used to estimate binary probability
- Model easy to estimate
- Extension to multiple outcomes