Introduction to the Logit Model (Logistic Regression)

data example : Credit_subset.csv

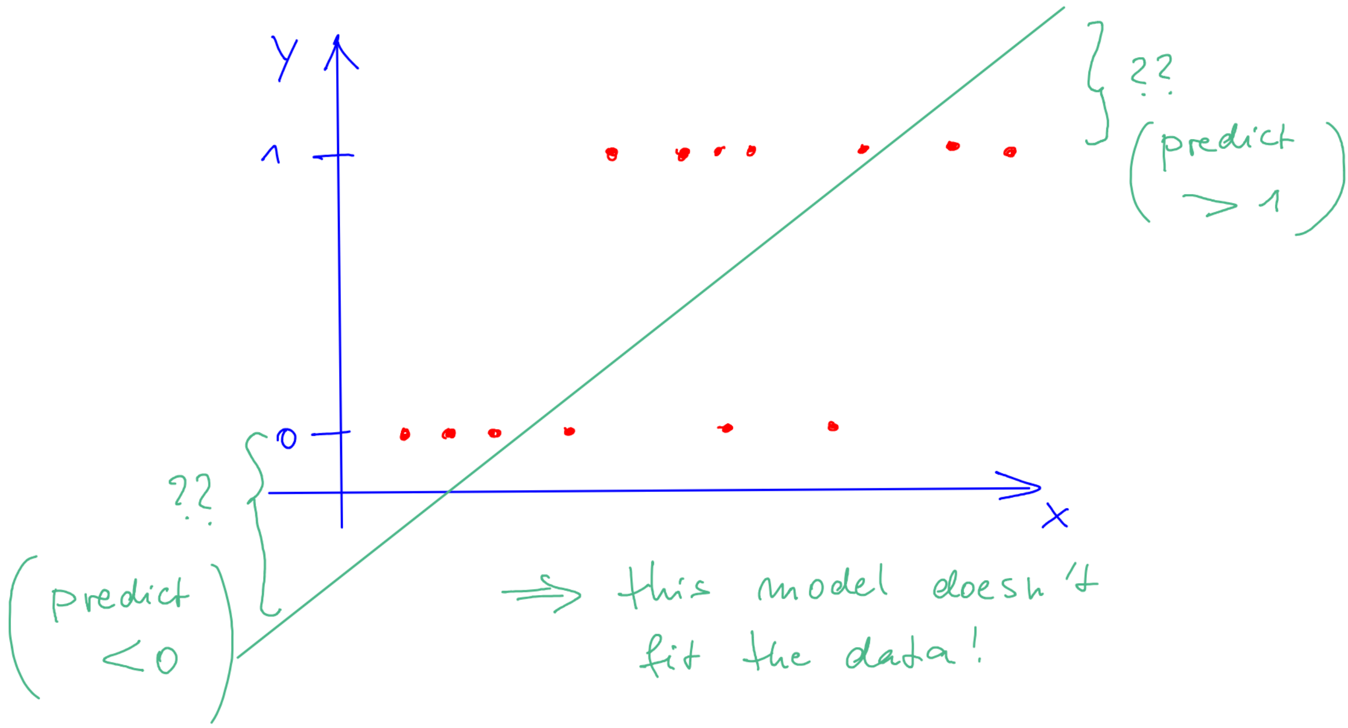
with $Y \in \{0, 1\}$

no credit default (0)

credit default (1) (loan is not paid)

plus : explanatory variables $X_1, \dots, X_p$

① first idea : estimate a linear model, for example with only one explanatory variable X (say amount of the loan)



$\Longrightarrow$ this model doesn't fit the data!

② idea : Conditional expectation

$$E(Y \mid X = x) = p_x = F(\beta^T x)$$

how to choose F ?

- if F would be a cdf
  then $p_x \in [0,1]$
  (or $p_x \in (0,1)$)
- if F would be strictly monotonously increasing, then as in the linear model:

  $\beta_j > 0$: expectation of Y is increasing if X_j increasing

  $\beta_j < 0$: expectation of Y is decreasing if X_j increasing
- also useful

  F is smooth $\Leftrightarrow$ continuously differentiable

  $\Longrightarrow$ it is easier to handle the optimization of the likelihood criterion

$\Longrightarrow$ $\boxed{\text{use a cdf for } F}$

$\Longrightarrow$ logit model $\hat{=}$ logistic regression

$$F(u) = \frac{1}{1+e^{-u}} = \frac{e^u}{1+e^u}$$

cdf of the standard logistic distribution

$\Rightarrow$ most easy to handle

$\Longrightarrow$ alternative : <u>probit model</u>

$$F(u) = \Phi(u)$$

cdf of the Gaussian distribution
$\hat{=}$ Standard normal / $N(0,1)$

<u>in R</u> : $\boxed{\text{glm}}$ (instead of lm)

generalized linear model

e.g. glm(y ~ x, family = binomial())

default : logit model

<u>or</u> :
glm(y ~ x, family = binomial(link = "logit"))

<u>or</u> :
glm(y ~ x, family = binomial(link = "probit"))

probit model

Almost everything works then
as in the linear model :

- coefficient estimator $\hat{\beta}$
- p-values when testing the coeff.
- degrees of freedom

...

Prediction

$$E(Y \mid X=x) = p_x \quad \text{and} \quad Y \sim B(1, p_x)$$

$$\Rightarrow \quad \underbrace{p_x = P(\underbrace{Y=1}_{\text{"Success"}} \mid X=x)}_{\text{"success" probability}}$$

$\Rightarrow$ using estimated coefficients $\hat{\beta}$

$$\hat{p}_x = F(\hat{\beta}^T x)$$

logit model : plogis
probit model : pnorm

in R : glm1 <- glm(....)

! predict(glm1, ...)
calculates $\underline{\hat{\beta}^T x}$ values
linear predictor

✓ predict(glm1, ... , type = "response")
calculates $\underline{F(\hat{\beta}^T x)}$ values
estimated probabilities

See ?predict.glm for help