Foundations of Data Science: Logistic regression

THE UNIVERSITY of EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Overview

- Principle of Logistic Regression

- Interpretation of Logistic Regression coefficients

- Multiple Logistic Regression

- Logistic Regression as a classifier

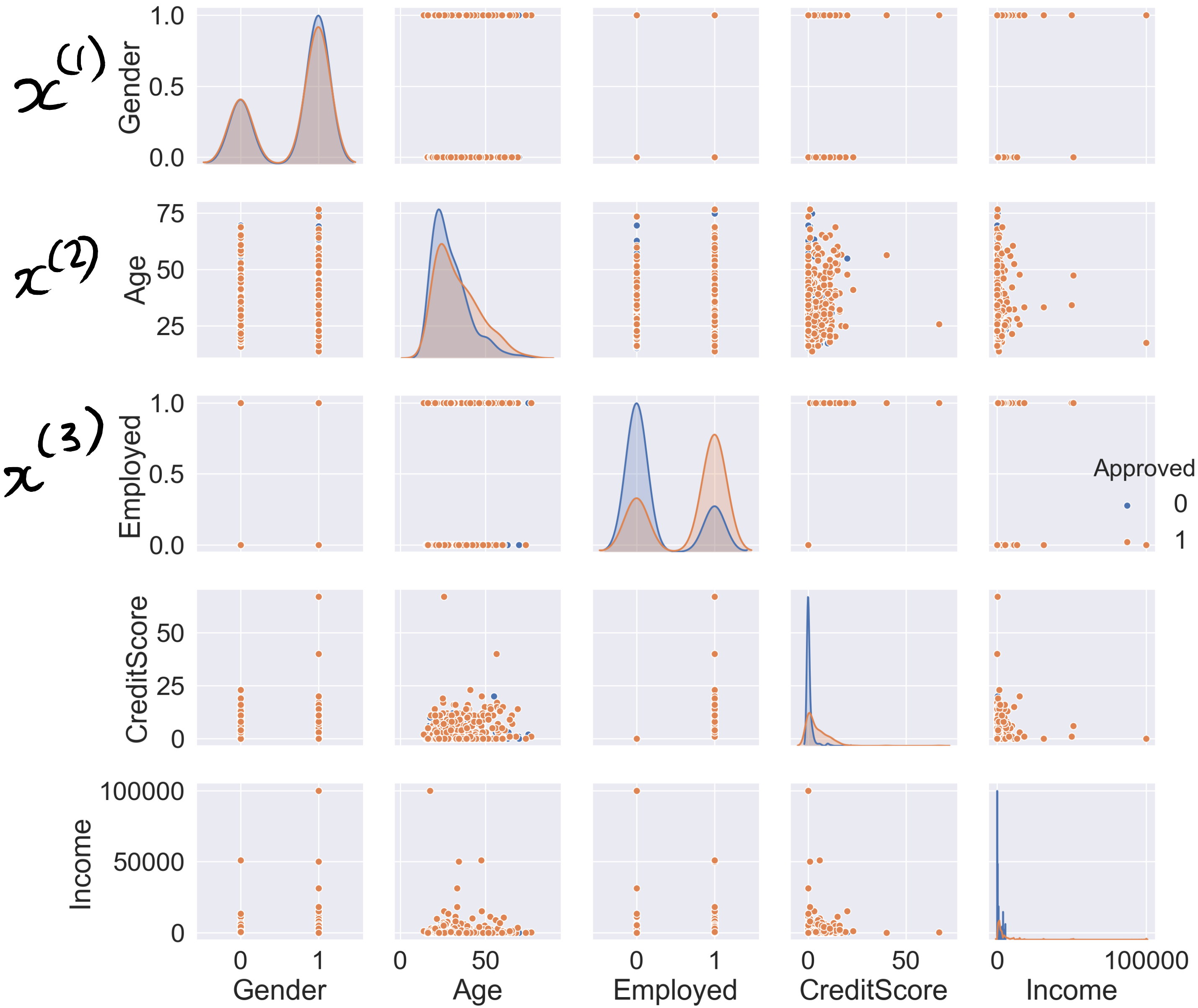
- (Maximum likelihood estimation of Logistic regression coefficients)

Foundations of Data Science: Logistic regression – Principle of logistic regression

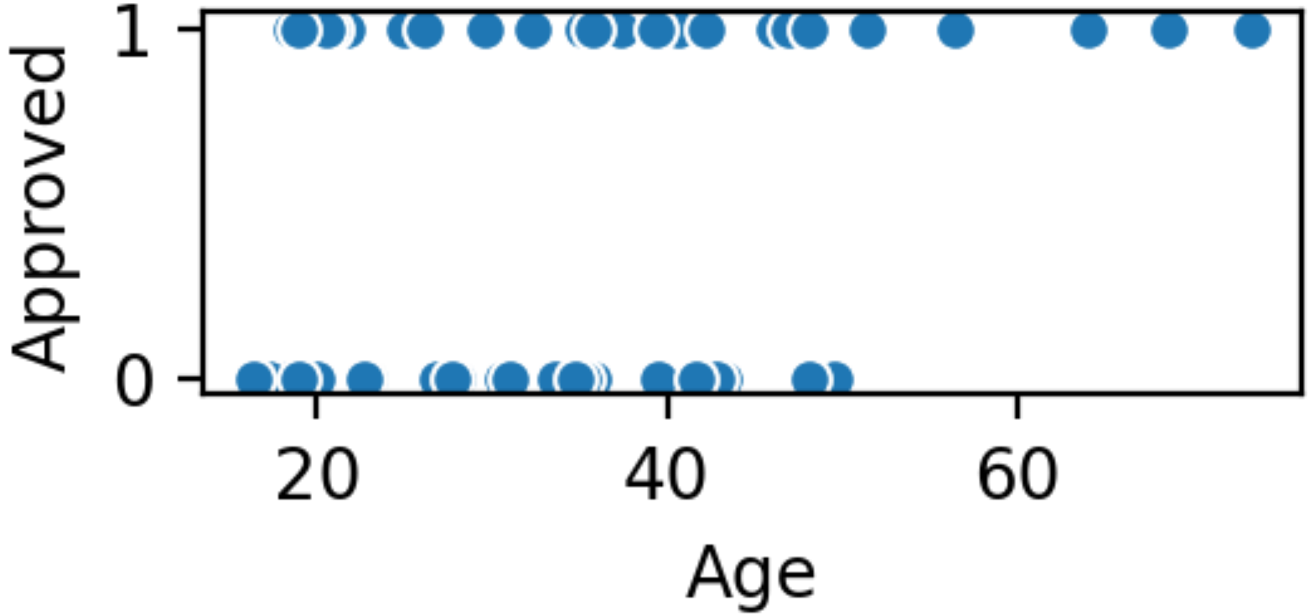
Supervised classification

Binary (or dichotomous) response variable: Credit

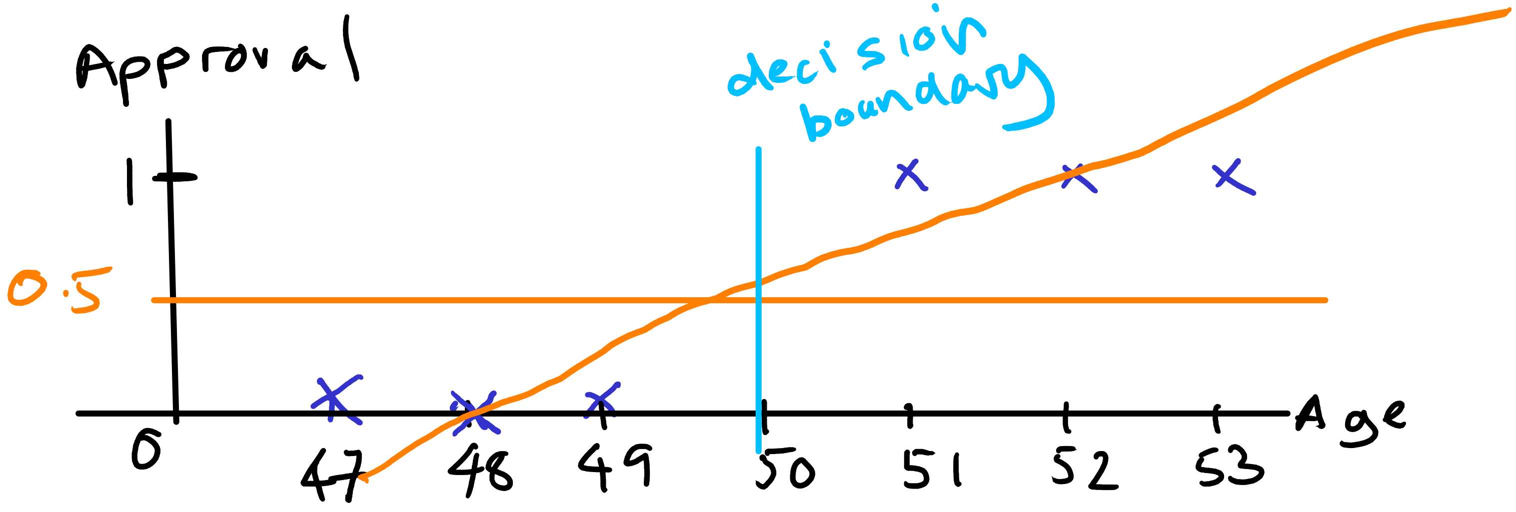
Approved
Not approved



Classification task on one continuous variable



Exercise :

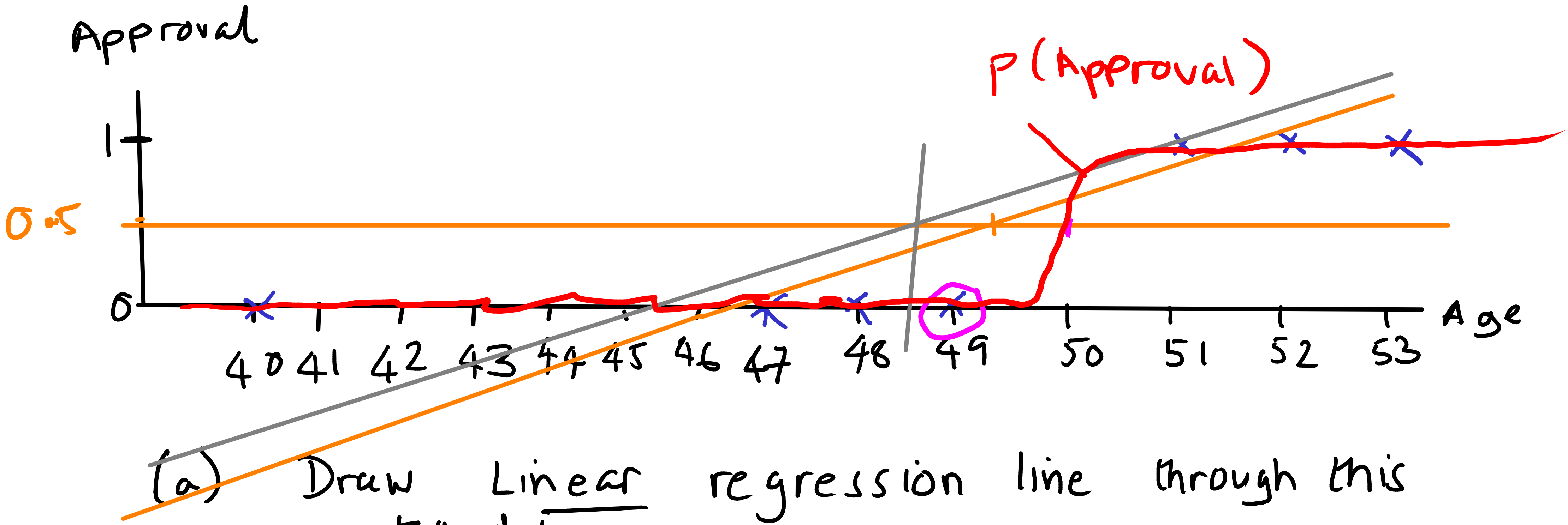


(a) Draw a Linear regression line through this toy data

(b) Convert the linear regression prediction to a predicted class label (0/1)

(c) Where is the decision boundary ?

(d) How many classification errors are there ?



(a) Draw Linear regression line through this toy data

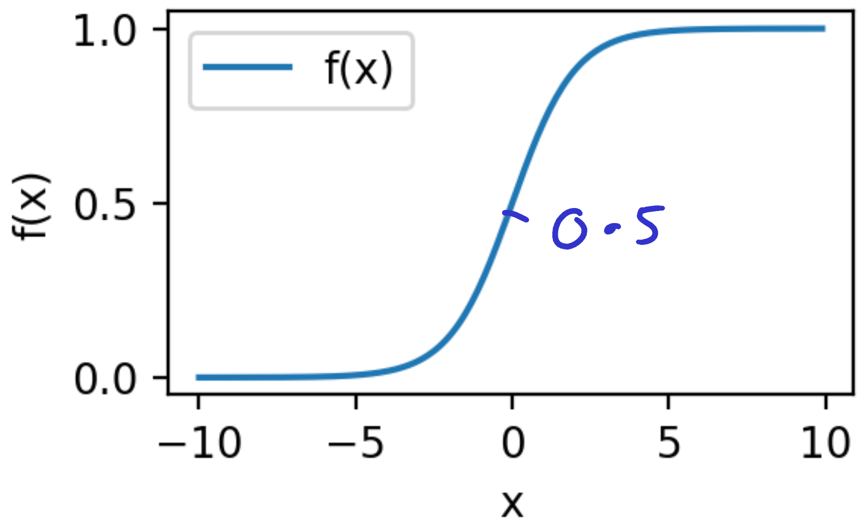
(b) Convert the linear regression prediction to a predicted class label (0/1)

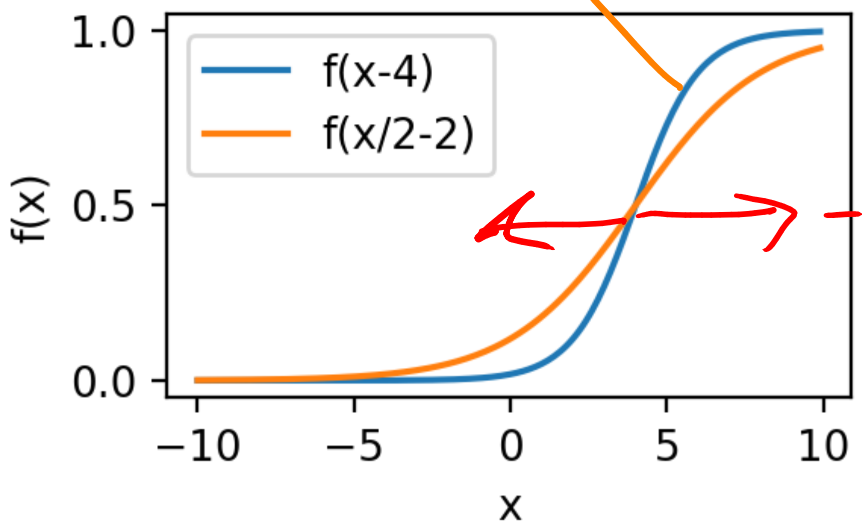
(c) Where is the decision boundary?

(d) How many classification errors are there?

Logistic function

β_1 controls slope





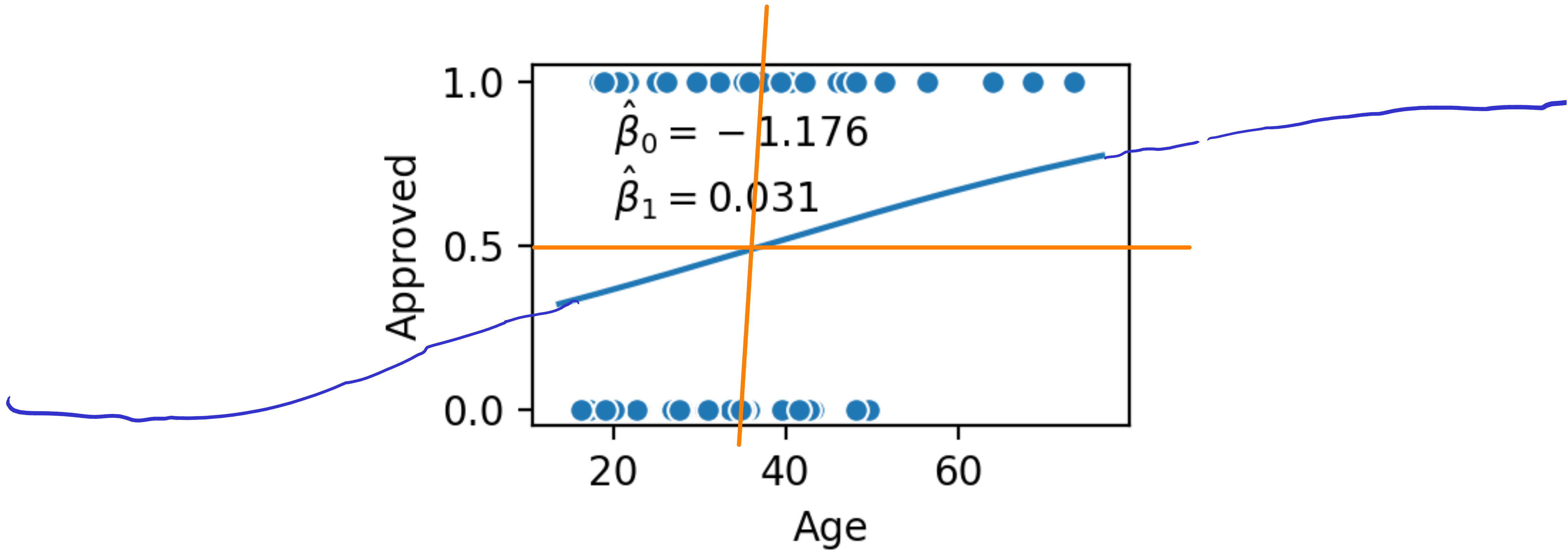
$-\frac{\beta_0}{\beta_1}$?

$$f(x) = \frac{e^{x}}{1+e^{x}} = \frac{1}{1+e^{-x}}$$

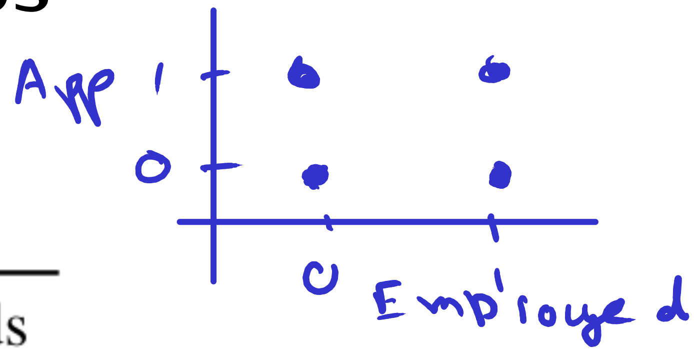
$$P(Y=1 \mid X=x) = f(\beta_0 + \beta_1 x) = \frac{1}{1+e^{-\beta_0-\beta_1 x}}$$

$\hat{\beta}_0$ $\hat{\beta}_0$

Application to continuous variable in credit example



Binary variables: odds and odds ratios



$P(Y = y \mid X = x)$

Employed	Approved	Not approved	Approval odds
0	0.25	0.75	0.34
1	0.71	0.29	2.42

$$OR(x) = \frac{2.42}{0.34} = 7.09$$

Effect size 609 %

$y \in \{$ "Not approved", "Approved" $\}$

$x \in \{$ "Not Emp.", "Emp." $\}$

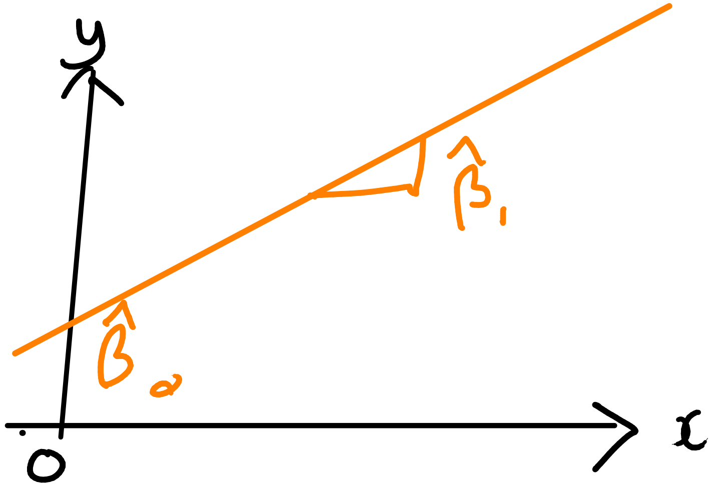
$$\text{Odds(Success)} = \frac{P(\text{Success})}{P(\text{Failure})} = \frac{P(\text{Success})}{1 - P(\text{Success})}$$

$$\text{Odds ratio } OR(x) = \frac{\text{Odds}(\text{Success} \mid x = \text{True})}{\text{Odds}(\text{Success}) \mid x = \text{False})}$$

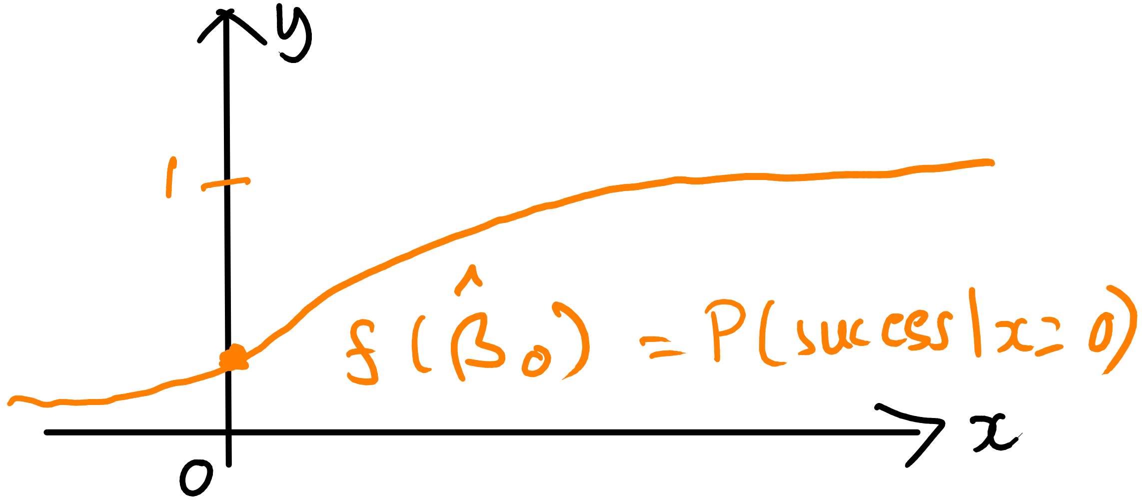
Foundations of Data Science: Logistic regression – Interpretation of logistic regression coefficients

Interpretation of $\hat{\beta}_0$

Lin reg



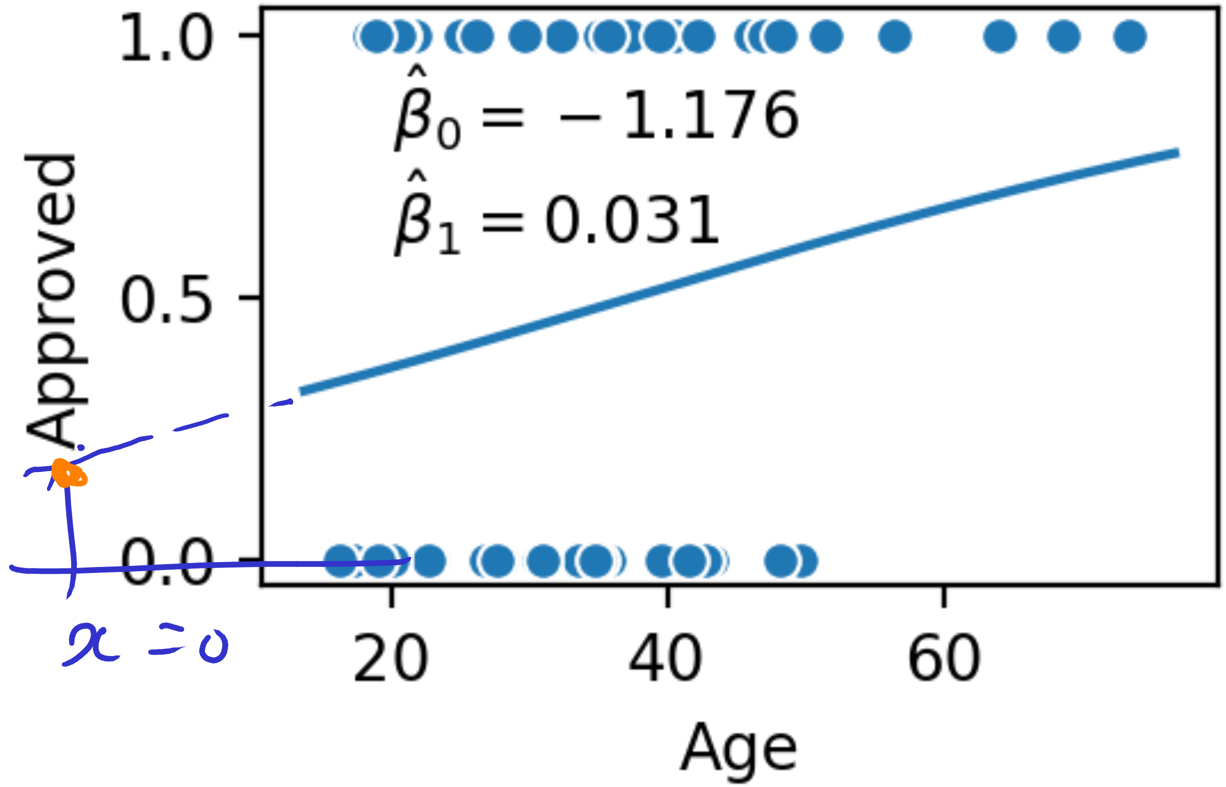
Log. reg



$$f(\hat{\beta}_0 + 0 \cdot \hat{\beta}_1) = f(\hat{\beta}_0)$$

$$= \frac{1}{1 + e^{-\beta_0}}$$

$$f(\hat{\beta}_0) = f(-1.176)$$
$$= 0.236$$



Log odds

$$\text{Log Odds(Success)} = \ln \frac{P(\text{Success})}{P(\text{Failure})} \quad (\ln = \log_e)$$

$$= \ln \frac{P(\text{Success})}{1 - P(\text{Success})} = \ln \frac{p}{1-p}$$

$$\boxed{\begin{array}{l} \text{Log odds} + 1 \\ \Rightarrow \text{odds} \times e \end{array}}$$

p	odds	Log odds
0.5	1	0
>0.5	>1	>0
<0.5	<1	<0
1	$\to \infty$	$\to \infty$
0	0	$\to -\infty$

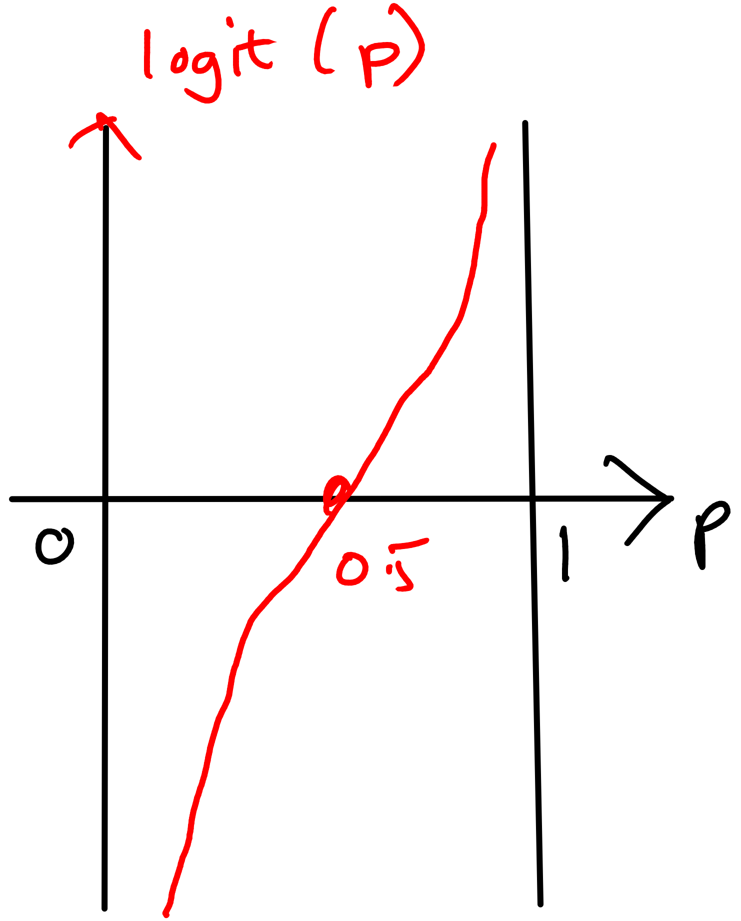
Logit scale

Log odds +1 $\Rightarrow$ odds increase by factor e

logistic unit = logit

$\hat{\beta}_0 = -1.176$ logits

$$\text{logit}(p) = \ln \frac{p}{1-p}$$



Logistic Regression in terms of log odds

Success $\quad P(Y=1 \mid x) = f(\beta_0 + \beta_1 x) = \dfrac{1}{1 + e^{-\beta_0 - \beta_1 x}}$ ①

Failure $\quad P(Y=0 \mid x) = 1 - f(\beta_0 + \beta_1 x) = 1 - \dfrac{1}{1 + e^{-\beta_0 - \beta_1 x}}$

$$= \frac{e^{-\beta_0 - \beta_1 x}}{1 + e^{-\beta_0 - \beta_1 x}}$$ ②

① / ②

Odds $\quad \dfrac{P(Y=1 \mid x)}{P(Y=0 \mid x)} = \dfrac{1}{e^{-\beta_0 - \beta_1 x}} = e^{\beta_0 + \beta_1 x}$

Log odds $\quad \ln \dfrac{P(Y=1 \mid x)}{P(Y=0 \mid x)} = \beta_0 + \beta_1 x = \text{logit}(P(Y=1 \mid x))$

Interpretation of $\hat{\beta}_1$

$$\text{Odds}(x) = e^{\hat{\beta}_0 + \hat{\beta}_1 x}$$

$$= e^{\hat{\beta}_0} e^{\hat{\beta}_1 x}$$

$x = \{0, 1\}$ $\quad \text{OR}(x) = \dfrac{\text{Odds}(1)}{\text{Odds}(0)}$

$$\text{OR}(x) = e^{\hat{\beta}_1}$$

$$\log \text{OR}(x) = \hat{\beta}_1$$

Credit e.g. $\text{OR}(\text{Age}) = e^{0.03} \simeq 1.03$

Foundations of Data Science: Logistic regression – Multiple logistic regression

Principle of multiple logistic regression

Predictor variables $x^{(1)}$: Age

$x^{(2)}$: Employment $= \{0, 1\}$

$$P(Y=1 \mid x^{(1)}, x^{(2)}, \dots)$$

$$= f(\beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \dots)$$

Multiple logistic regression applied to the credit example

	Variable	Coefficient	Odds or OR	
$\hat{\beta}_0$	Intercept	–1.969	0.140	← odds
$\hat{\beta}_1$	Age	0.029	1.030	← OR
$\hat{\beta}_2$	Employed	1.881	6.562	← OR

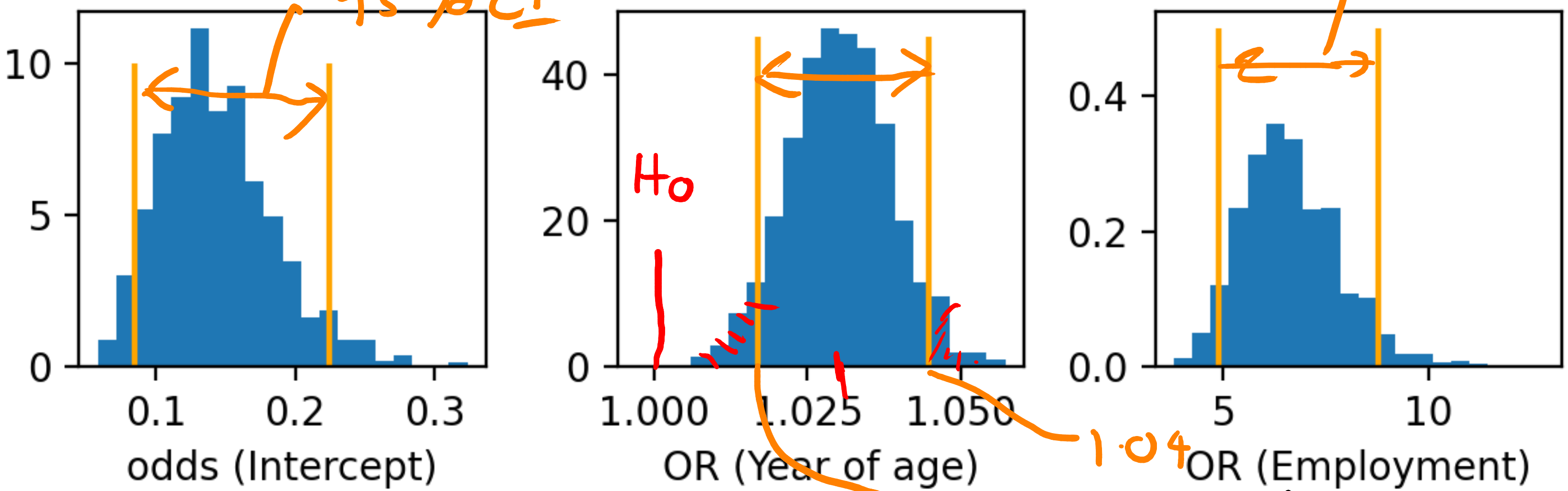
log odd
logits

$e^{+\hat{\beta}}$

Bootstrap confidence intervals

$B = 1000$ $\quad n = 653$

95% CI $\quad$ 95% CI $\quad$ 95% CI



$e^{\hat{\beta}_0}$ $\quad$ $e^{\hat{\beta}_1}$ $\quad$ $e^{\hat{\beta}_2}$

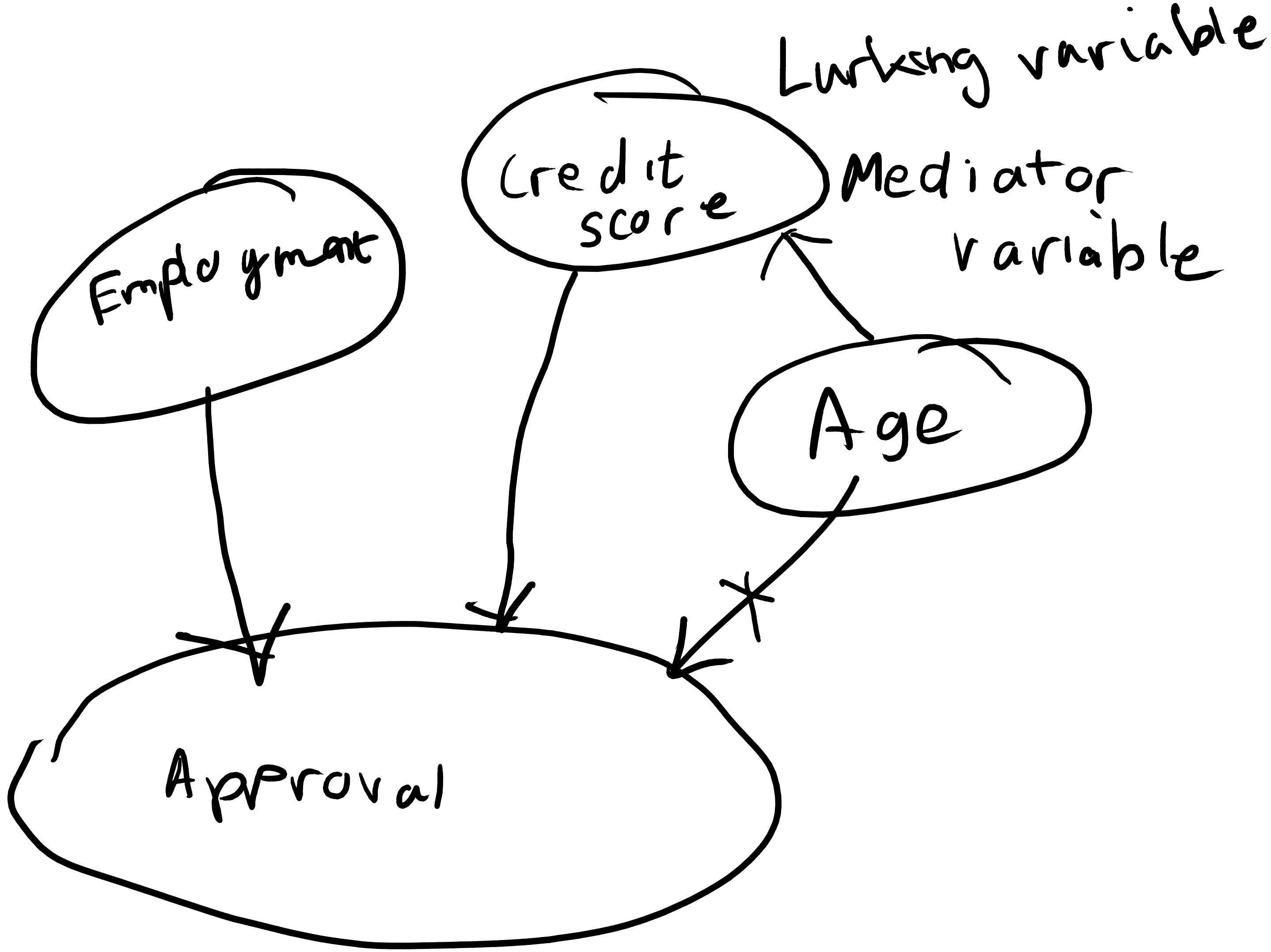
Does age affect credit approval?

H_0: age does not affect credit approval $\Rightarrow e^{\hat{\beta}_1} = 1$

H_a: " " affect credit approval in some way.

Discussion question

Can you think of any problems in the reasoning that we've used to suggest age and credit approval are related?


Lurking variable
Credit score
Mediator variable
Employment
Age
Approval

Foundations of Data Science: Logistic regression – The logistic regression classifier

Converting logistic regression to a classifier

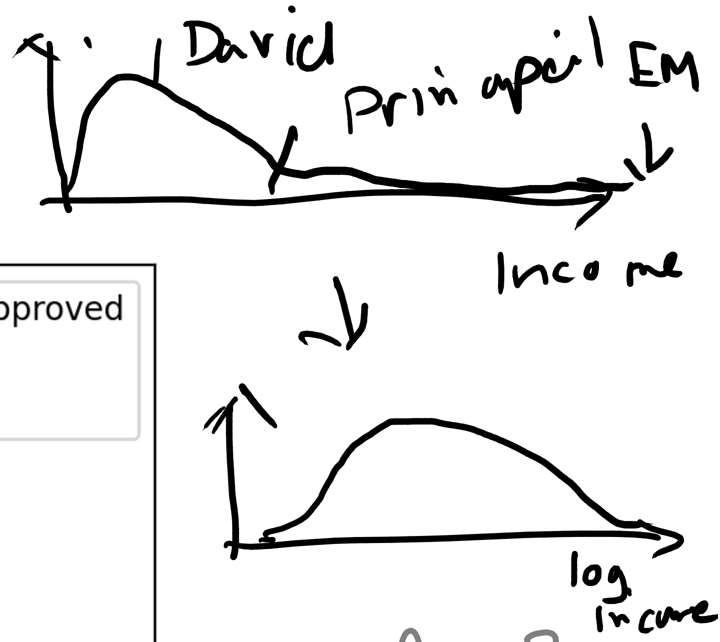
- Fit logistic regression model
- Set threshold c in terms of log odds and apply to predicted log odds:

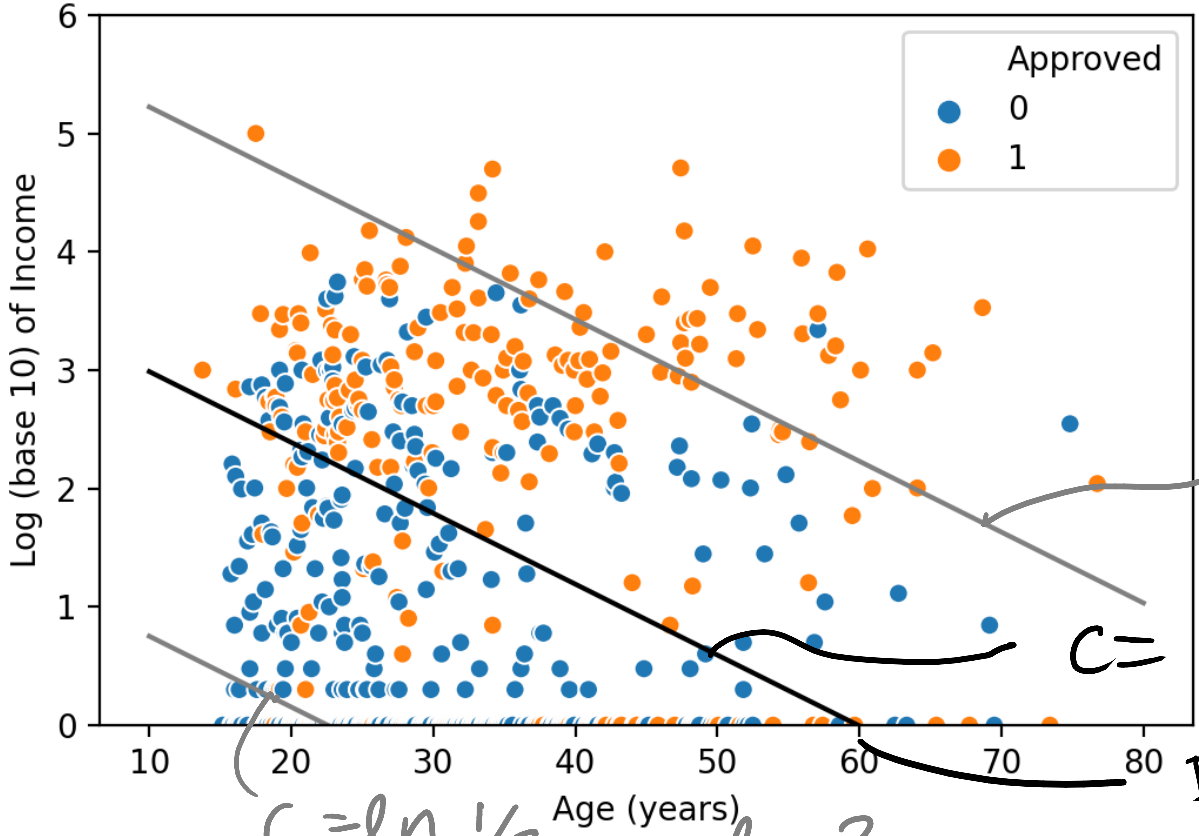
$$\hat{\beta}_0 + \hat{\beta}_1 x^{(1)} + \hat{\beta}_2 x^{(2)} + \ldots \geq c \Rightarrow \hat{y} = 1$$

$$\hat{\beta}_0 + \hat{\beta}_1 x^{(1)} + \hat{\beta}_2 x^{(2)} + \ldots < c \Rightarrow \hat{y} = 0$$

$c = 0 \Rightarrow$ odds of $1 \Rightarrow p = 0.5$

Decision boundary





$c = \ln 3$

odds e^3

$p = \text{logit}(\ln 3)$

$C = 0$

Decision boundary

$c = \ln 1/3 = -\ln 3$

$$\beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \dots > c \Rightarrow \hat{y} = 1$$

$$\beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \dots \le c \Rightarrow \hat{y} = 0$$

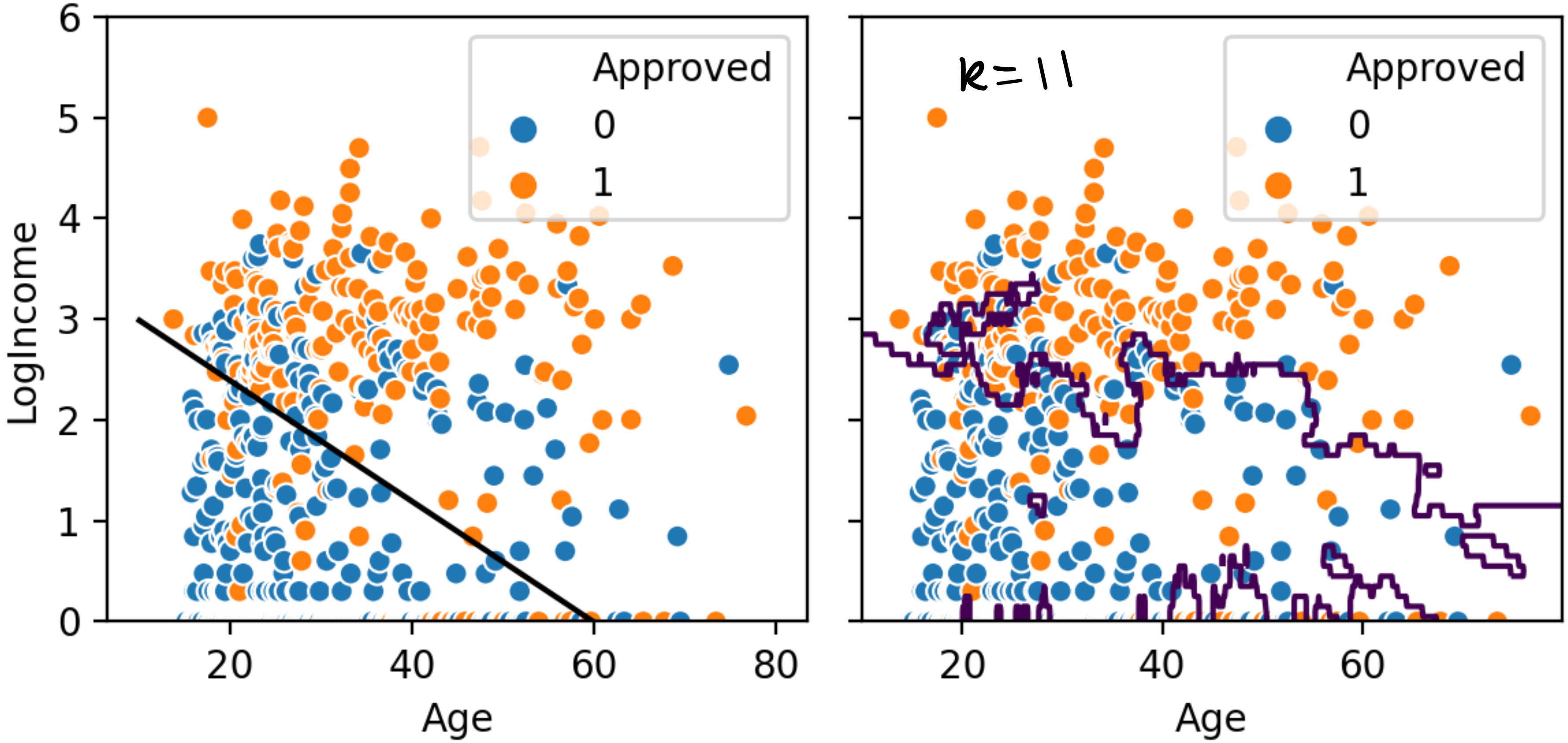
Ethics: logistic regression can be transparent

Credit scoring system:

- If you are in employment you score 1.625, if not you score 0
- Multiply your age by 0.029 and add the result to your score
- Round your income to the nearest 1000.
  Multiply the number of zeros in this figure by 0.320
  and add the result to your score
- If you scored more than 2.246, your credit will be approved

Cf. "Promote Values of Transparency, Autonomy and Trustworthiness" (Vallor, 2018)

Logistic regression versus k-NN



Decision boundary, flexibility/over-fitting, transparency

Standardised input variables

Summary

- Interpret $\hat{\beta}_0$ and $\hat{\beta}_1$ in terms of log odds

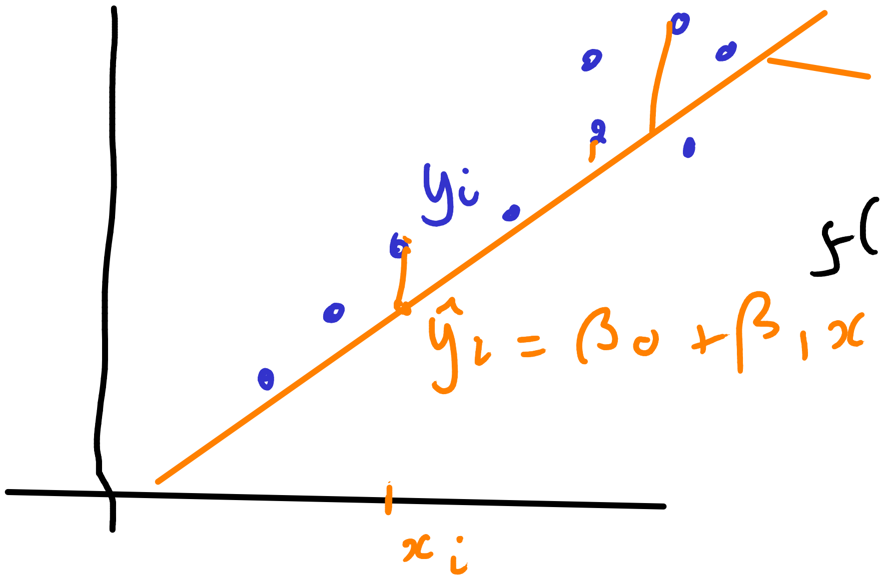
- Extend logistic regression to multiple variables

- Use logistic regression as a classifier

- Not covered (yet): derivation of logistic regression from principle of max likelihood

Foundations of Data Science: Logistic regression – Maximum likelihood estimation of logistic regression coefficients

Principle of Maximum Likelihood



$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x)^2$$

Optimisation:

$\beta_0 =$ $\qquad$ $\beta_1 =$

Adjust coefficients so as to maximise the likelihood of the data.

$\Rightarrow$ Expression for max. likelihood

Optimise w.r.t $\beta_0, \beta_1, \ldots$

Likelihood of one point

$$P(Y_i = 1 \mid X_i = x_i) = f(\beta_0 + \beta_1 x_i) \tag{1}$$

$$P(Y_i = 0 \mid X_i = x_i) = 1 - f(\beta_0 + \beta_1 x_i) \tag{2}$$

$$\Rightarrow P(Y_i = y_i \mid X_i = x_i) =$$

$$\underbrace{y_i f(\beta_0 + \beta_1 x_i)}_{1} + \underbrace{(1 - y_i)(1 - f(\beta_0 + \beta_1 x_i))}_{2}$$

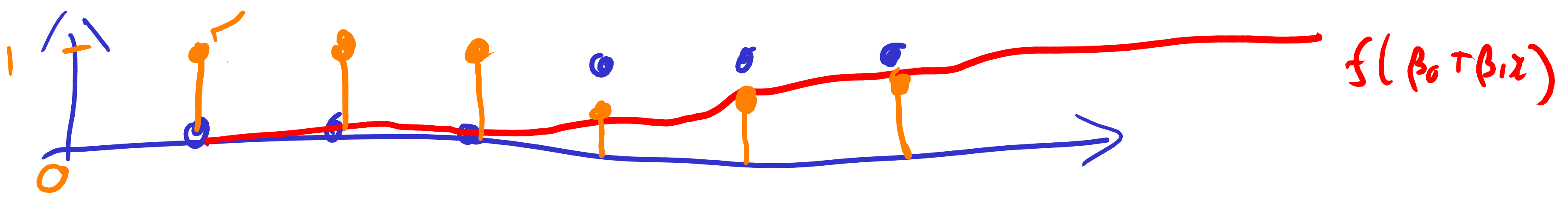
Likelihood of data given model

Assumption: responses are independent, given value of predictor variables

$$P(Y = y \mid X = \underline{x}) = P(Y = y_1 \mid X = x_1)\, P(Y = y_2 \mid X = x_2) \ldots$$

$$= \prod_{i=1}^{n} P(Y = y_i \mid X = x_i)$$

$$= \prod_{i=1}^{n} \left\{ y_i f(\beta_0 + \beta_1 x_i) + (1 - y_i)(1 - f(\beta_0 + \beta_1 x_i)) \right\}$$



Likelihood of data given model

$$\ln ab = \ln a + \ln b$$

$$\ln \prod_{i=1}^{n} a_i = \sum_{i=1}^{n} \ln a_i$$

Log likelihood:

$$\ln P(Y = y \mid X = \underline{x}) =$$

$$\sum_{i=1}^{n} \ln \left\{ y_i f(\beta_0 + \beta_1 x_i) + (1 - y_i)(1 - f(\beta_0 + \beta_1 x_i)) \right\}$$

https://covid.joinzoe.com/