Inf2 - Foundations of Data Science: Multiple logistic regression for explanation and prediction

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Announcements

- Week 4 workshop - we'll look at the paper that we'll be refer to in the exam

- Uses concepts from today's lecture!

- Solutions for Week 3 Workshop now available

- Solutions for this Week 4 Workshop will be available later in the week

- Badges on order!

Where we're at in the Maximum Likelihood Principle and Regression

Week 4: Logistic regression

Week 5: The maximum likelihood principle, and how we can use it to derive linear, logistic and other types of regression

Today

- Recap

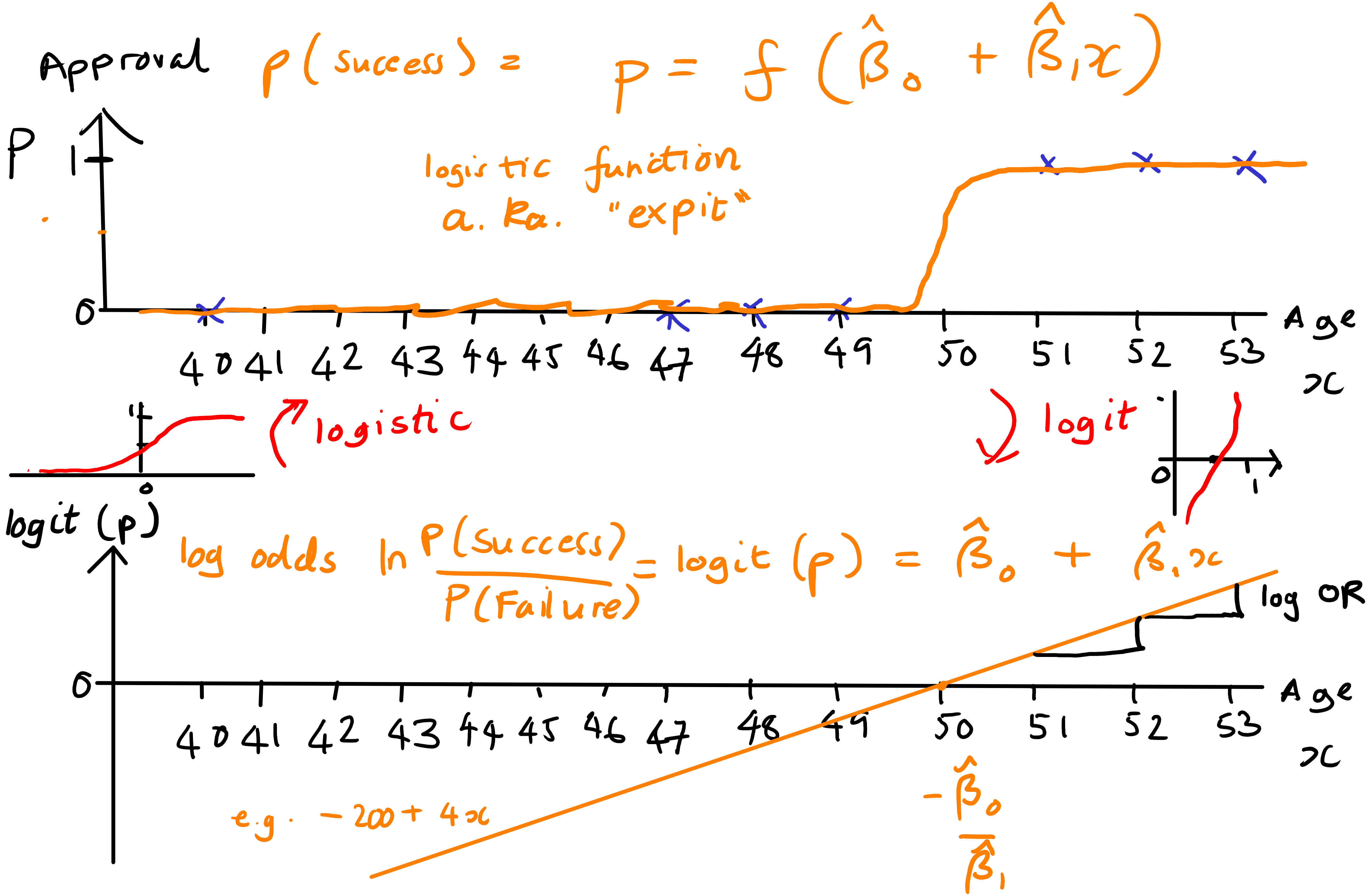
- Multiple Logistic Regression

- Confidence intervals on coefficients

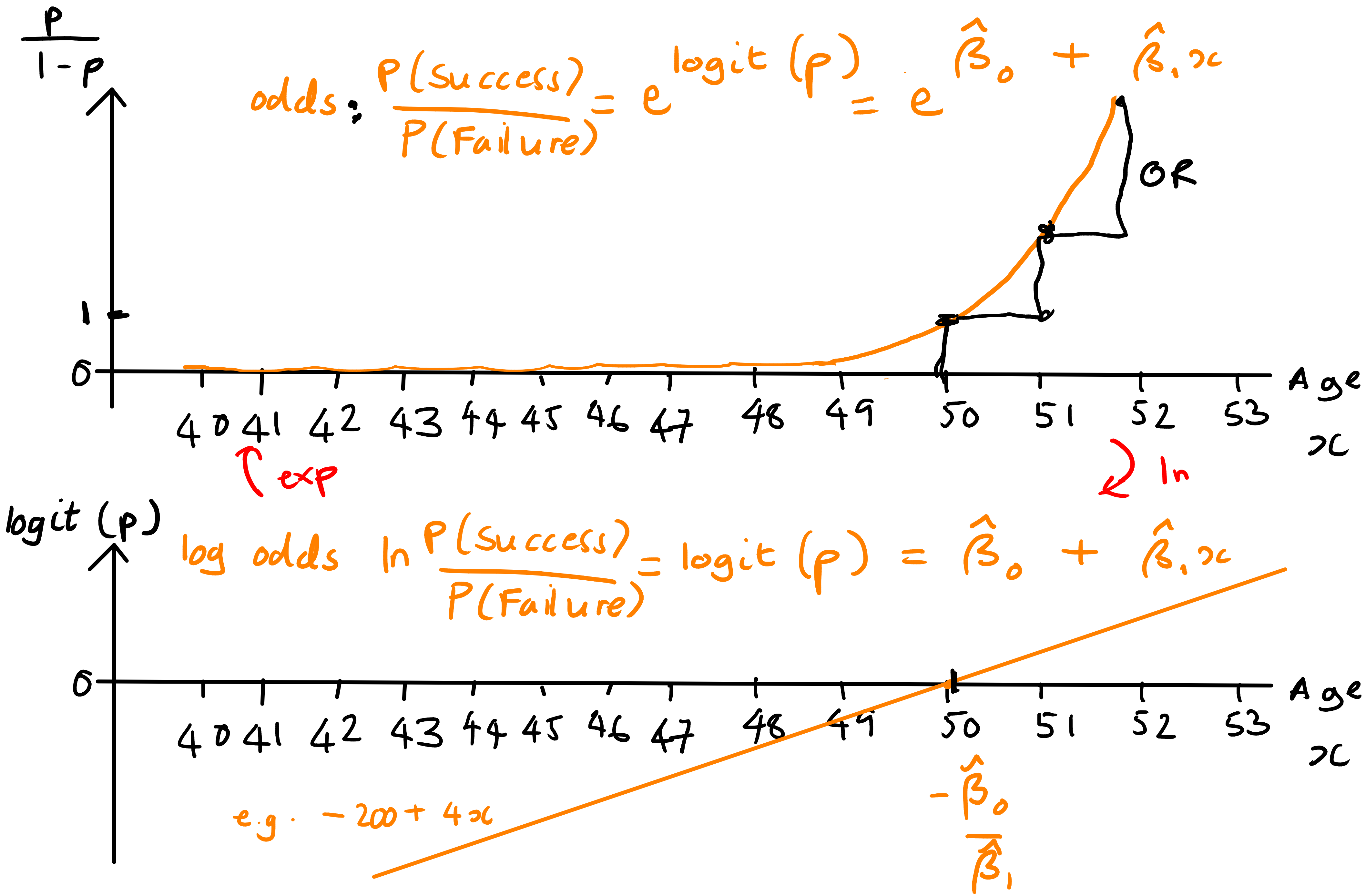
- Machine learing: Logistic Regression as a classifier

- Ethics of logistic regression

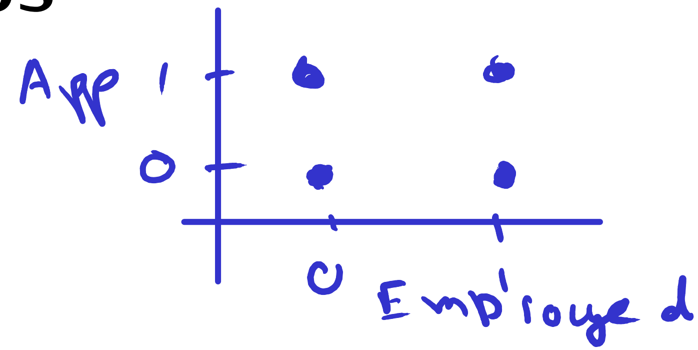
Probability and log odds views of logistic regression



Odds and log odds views of logistic regression



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

p $\qquad$ p $\qquad$ $\frac{p}{1-p}$

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

Effect size 609 %

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

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

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

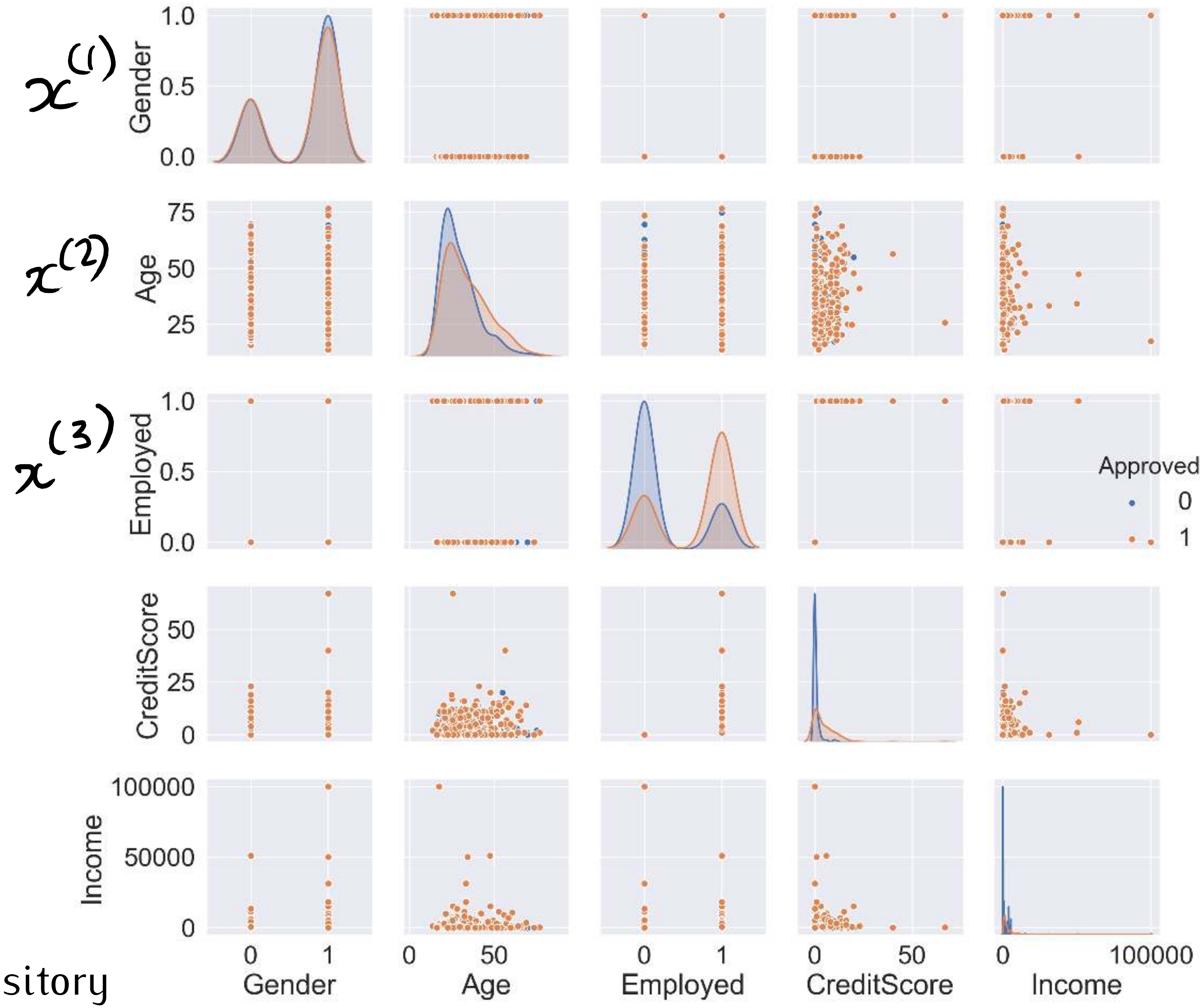
$$\text{Odds ratio } OR(x) = \frac{\text{Odds(Success} \mid x = \text{True})}{\text{Odds(Success)} \mid x = \text{False})}$$

Inf2 - Foundations of Data Science: Multiple logistic regression

Supervised classification

Binary (or dichotomous) response variable: Credit

Approved
Not approved



UCI Machine Learning Repository
https://archive.ics.uci.edu/dataset/27/credit+approval

Principle of multiple logistic regression

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

$x^{(2)}$: Employment

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

$$= f(\hat{\beta}_0 + \hat{\beta}_1 x^{(1)} + \hat{\beta}_2 x^{(2)} + \ldots)$$

↑ logistic

Multiple logistic regression applied to the credit example

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

$e^{\hat{\beta}_0}$ ← odds

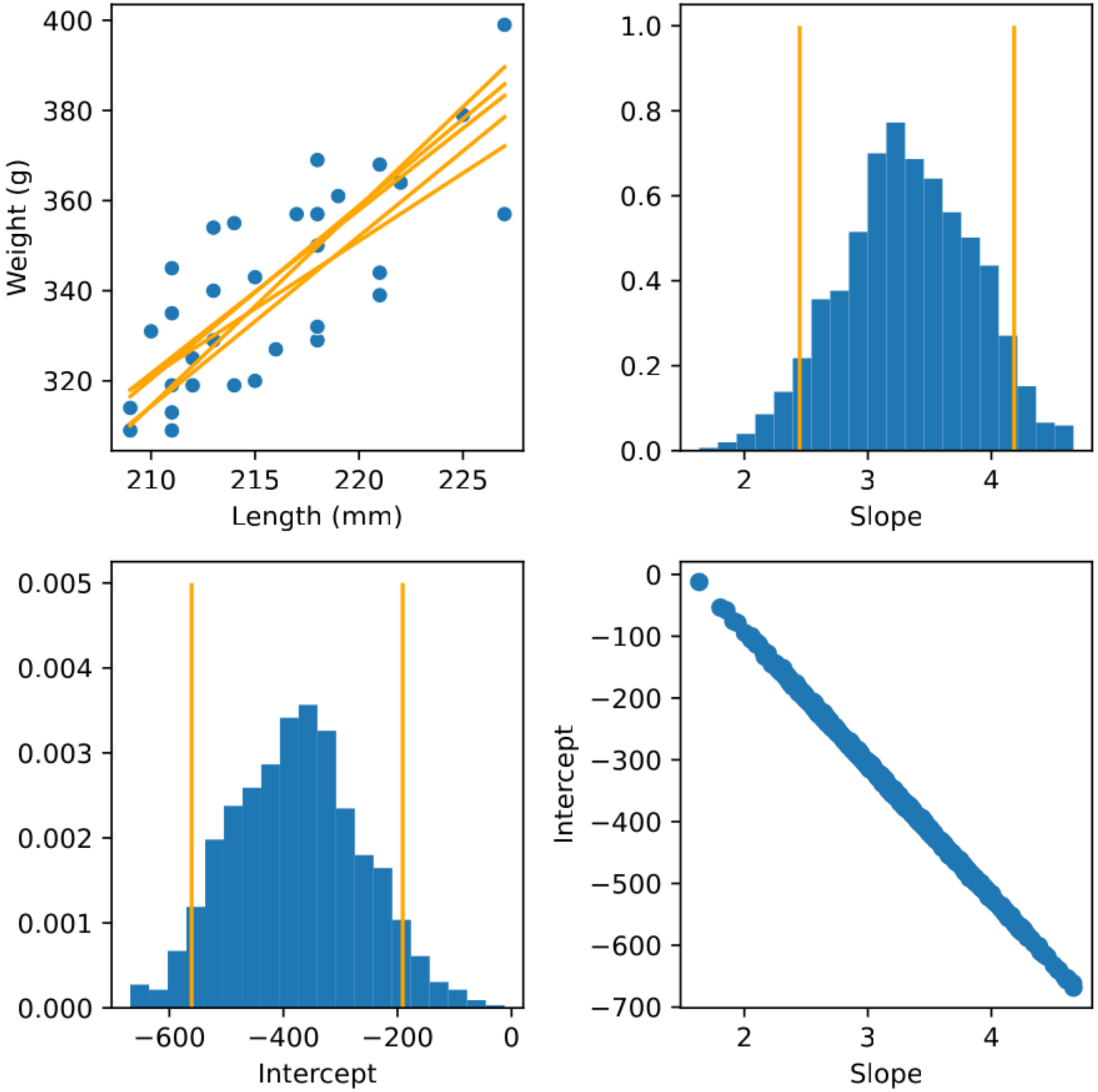
← OR $e^{\hat{\beta}_1}$

← OR $e^{\hat{\beta}_2}$

↑ log odds

logits

Boostrap confidence intervals for regression coefficients

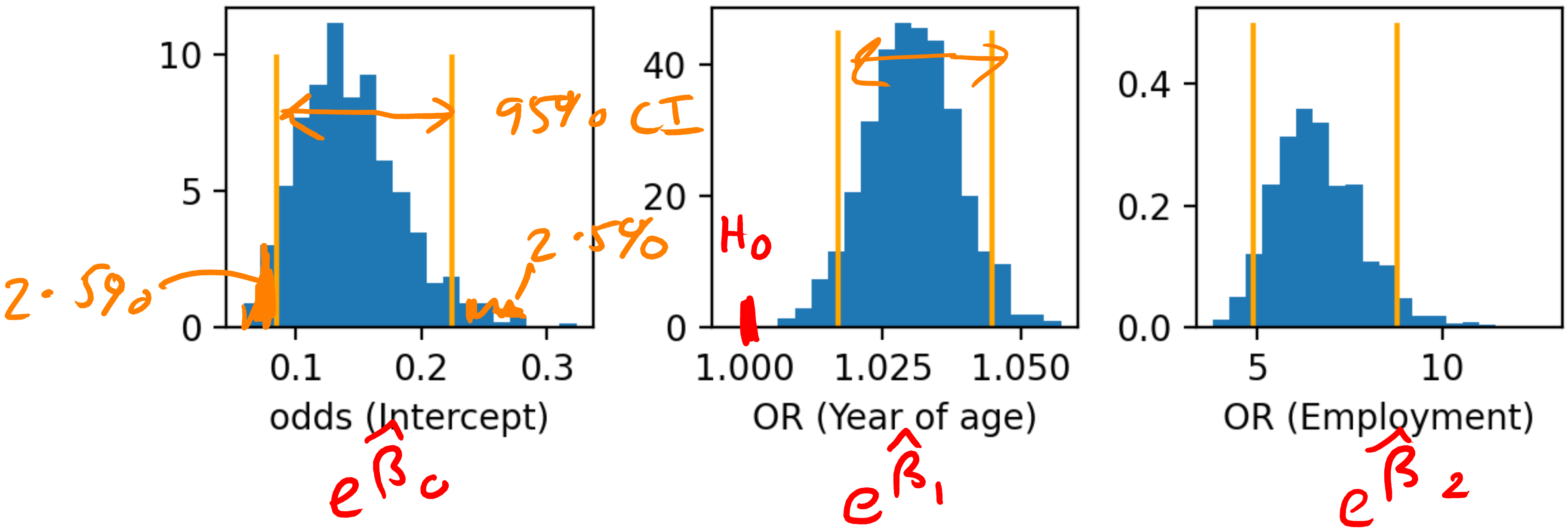


Demo

Code this for Logistic Regression in the lab!

Bootstrap confidence intervals

$B = 1000$ $n = 653$



Does age affect credit approval?

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

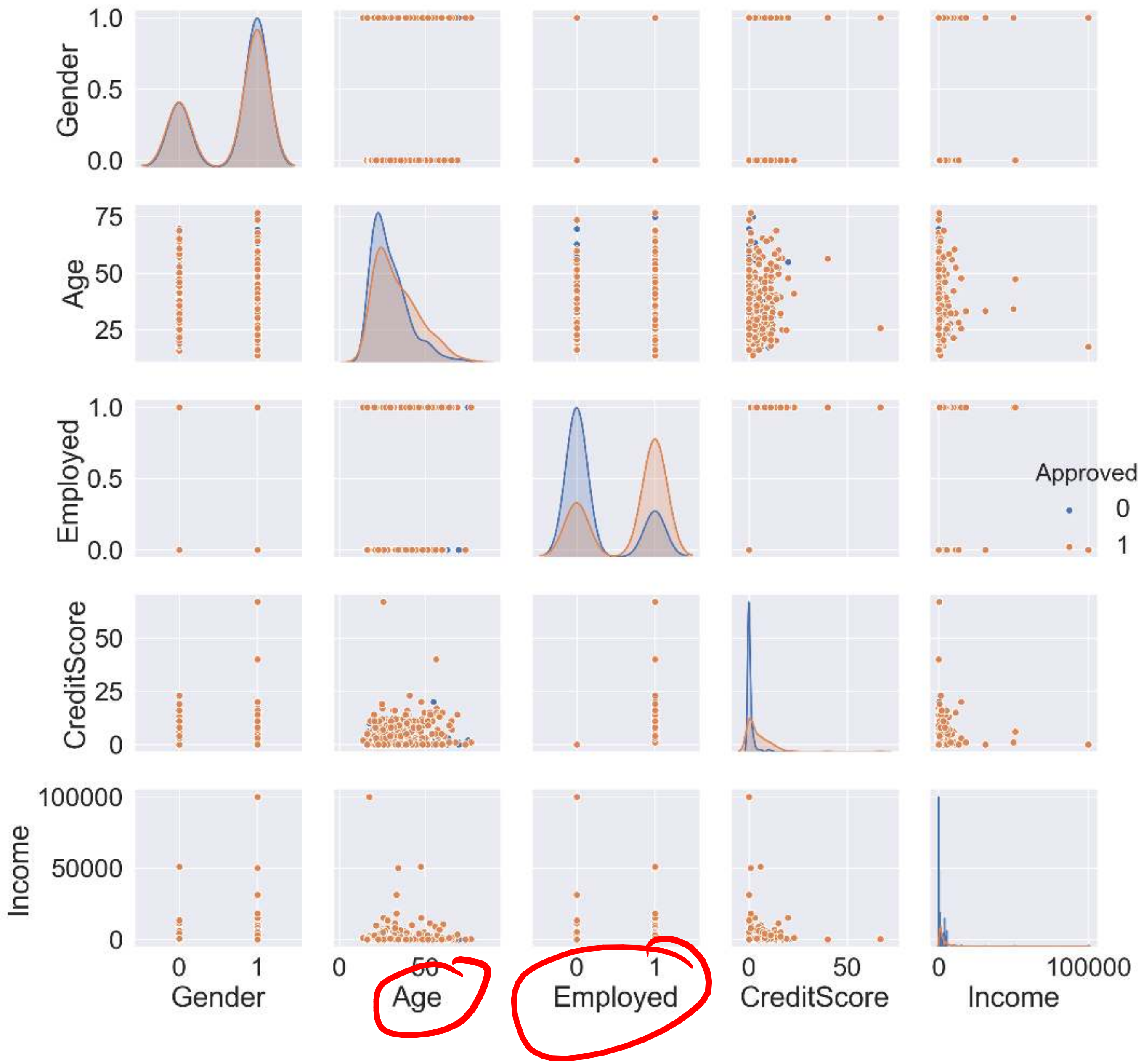
H_a: age does affect credit approval

Discussion question

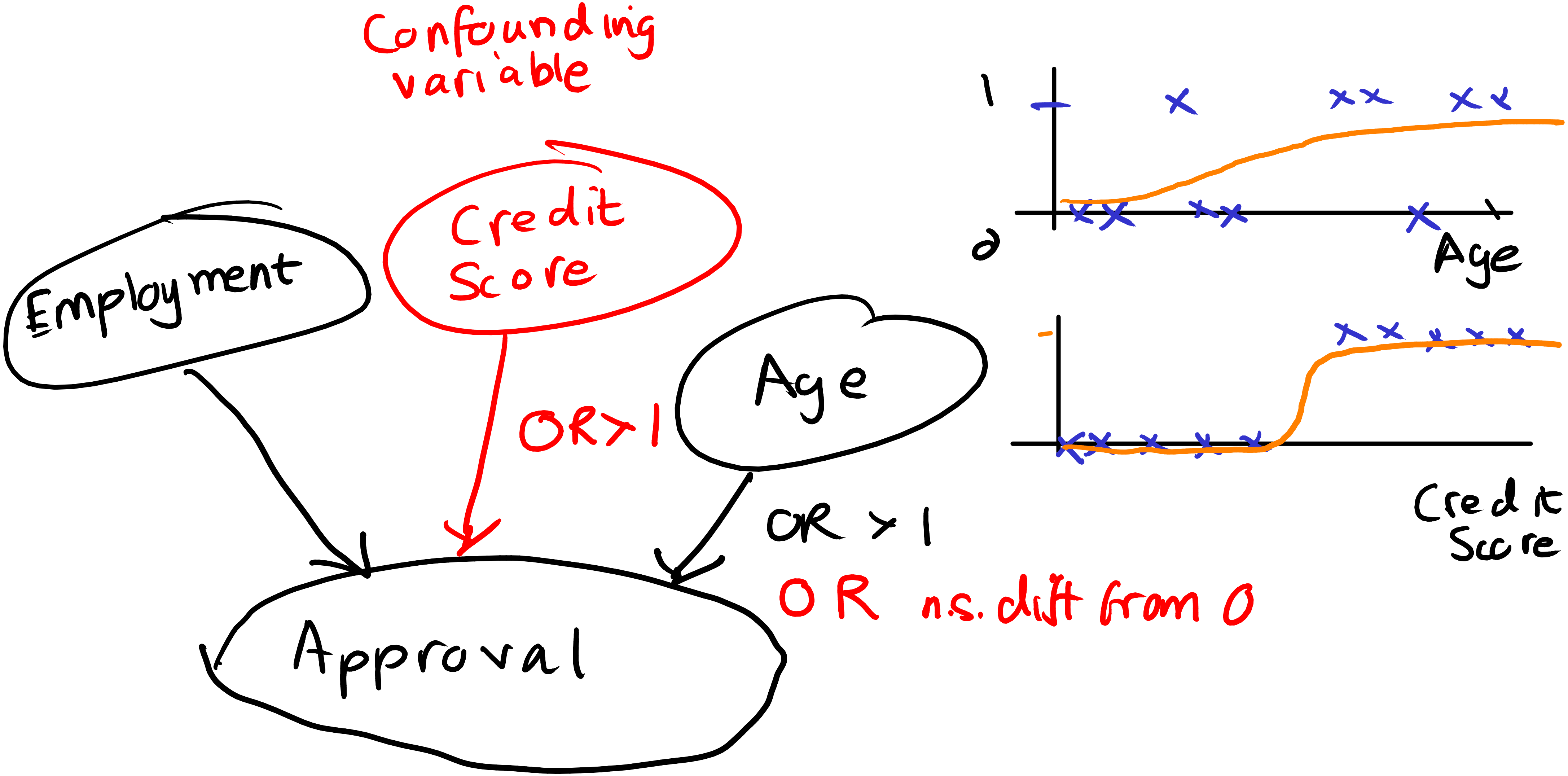
Our analysis so far shows that age and credit approval are related.
So all other things being equal, a 20 year old is less likely to have credit approved than a 50 year old.

Do we believe this yet?

What further analysis should we do?



Explanation - "controlling for", "adjusting for"



This week's lab

Multiple logistic regression on fuller set of variables

Using Logsitic Regression as a Machine Learning algorithm

Controlling for variables in the news: 5 February 2025

BBC | For you | Home | News | Sport | Weather | iPlayer | Sc

NEWS

Home | InDepth | Israel-Gaza war | War in Ukraine | Climate | UK | World | Business | Politics | Culture

Family & Education | Young Reporter

School phone bans don't boost grades or wellbeing, study suggests

Alice Evans
BBC News

THE LANCET *Regional Health*
Europe

This journal | Journals | Publish | Clinical | Global health | Multimedia | Events | About

ARTICLES • Online first, 101211, February 04, 2025 • *Open Access*

School phone policies and their association with mental wellbeing, phone use, and social media use (SMART Schools): a cross-sectional observational study

Victoria A. Goodyear [a,b] • Amie Randhawa [a,b] • Péymane Adab [c] • Hareth Al-Janabi [b,c] • Sally Fenton [a,d] • Kirsty Jones [e]
• et al. Show more

Inf2 - Foundations of Data Science: The logistic regression classifier

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Converting logistic regression to a classifier

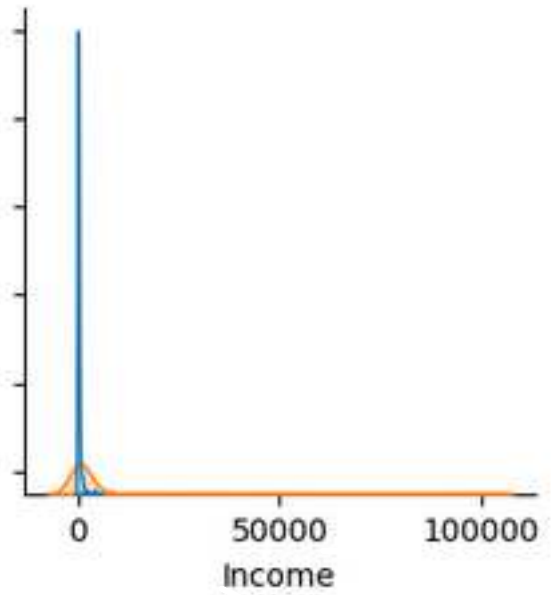
- Fit logistic regression model to data
- Set threshold c in terms of log odds and apply to predicted log odds

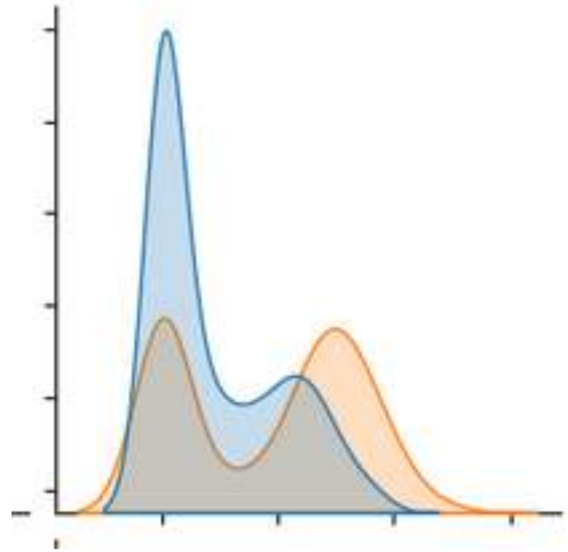
$$\hat{\beta}_0 + \hat{\beta}_1 x^{(1)} + \hat{\beta}_2 x^{(2)} + \ldots \geq c \Rightarrow \hat{y} = 1$$

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

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

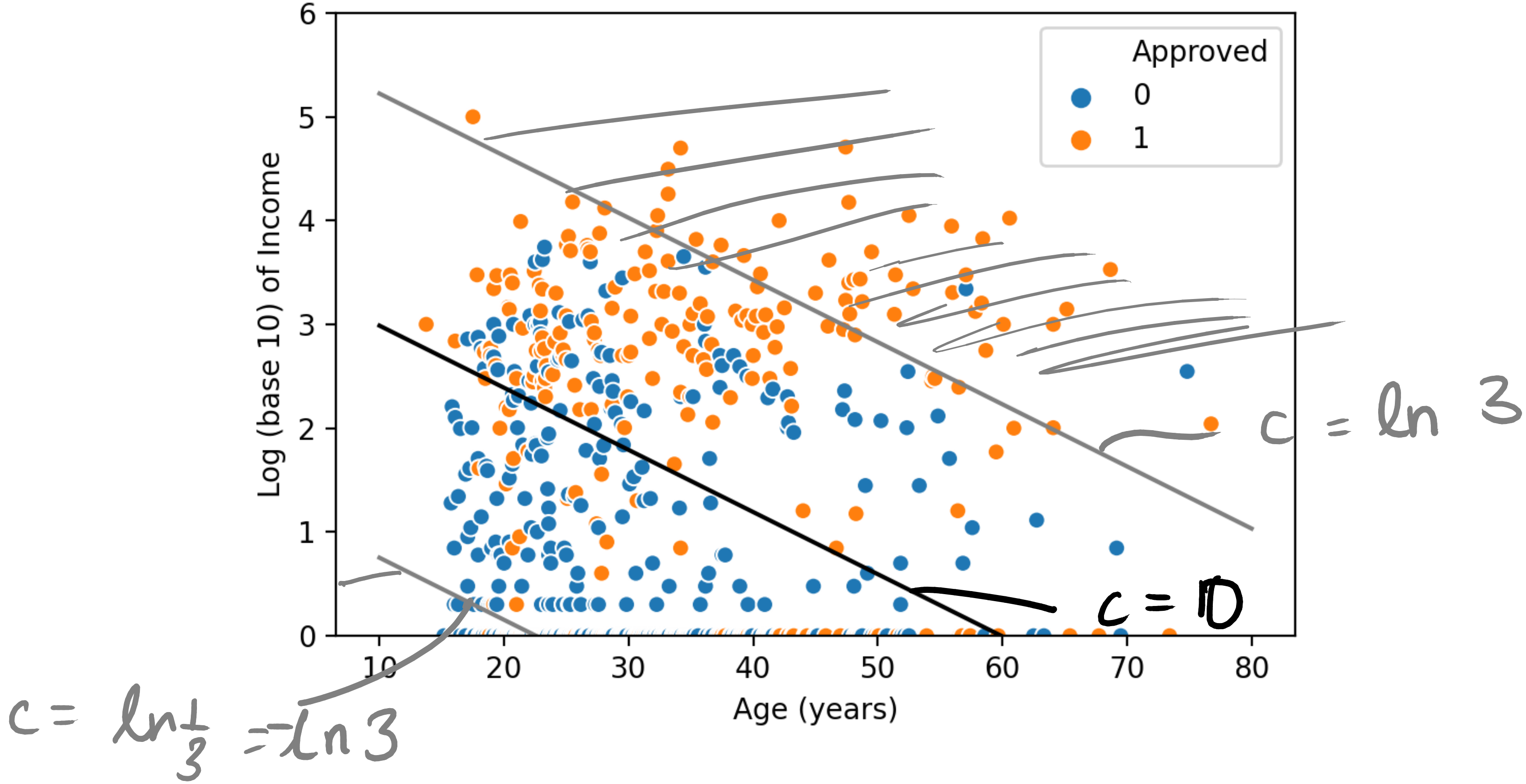
Machine learning trick: make marginal distributions more normal





Log income

Decision boundary



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

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

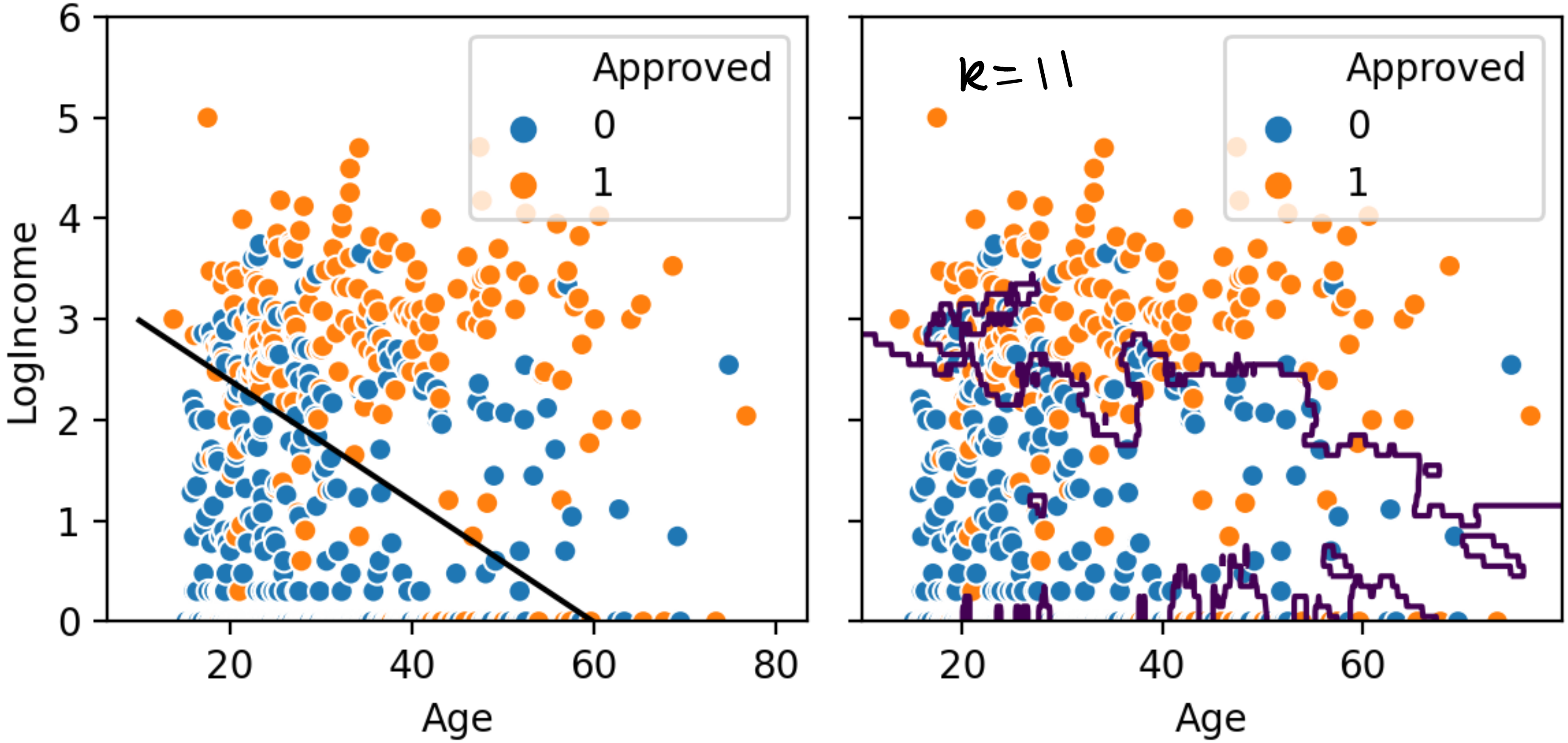
Ethics: logistic regression can be transparent

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)

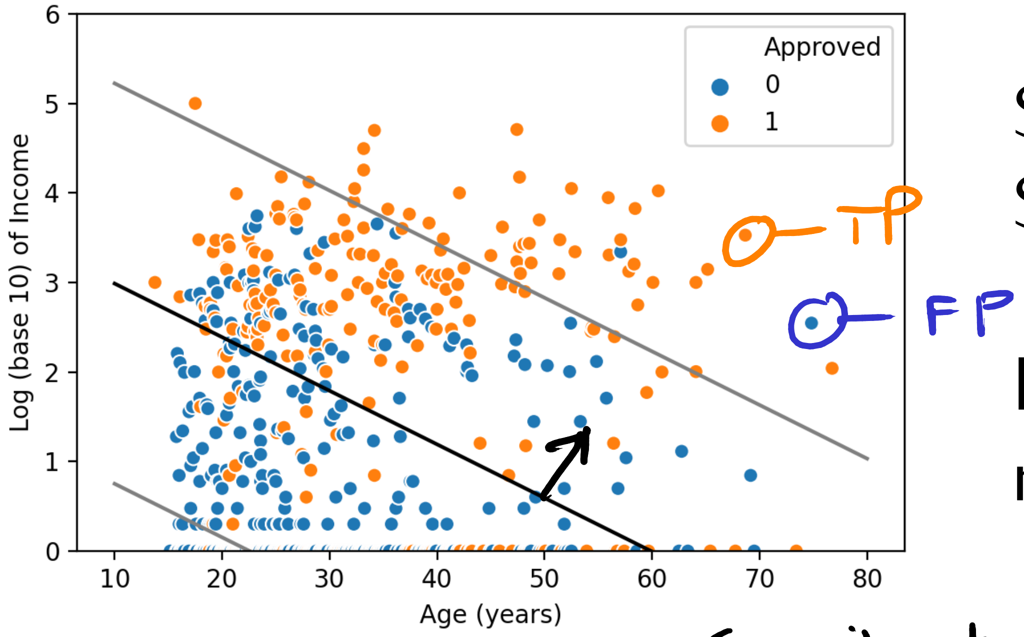
Logistic regression versus k-NN



Decision boundary, flexibility/over-fitting, transparency

Standardised input variables

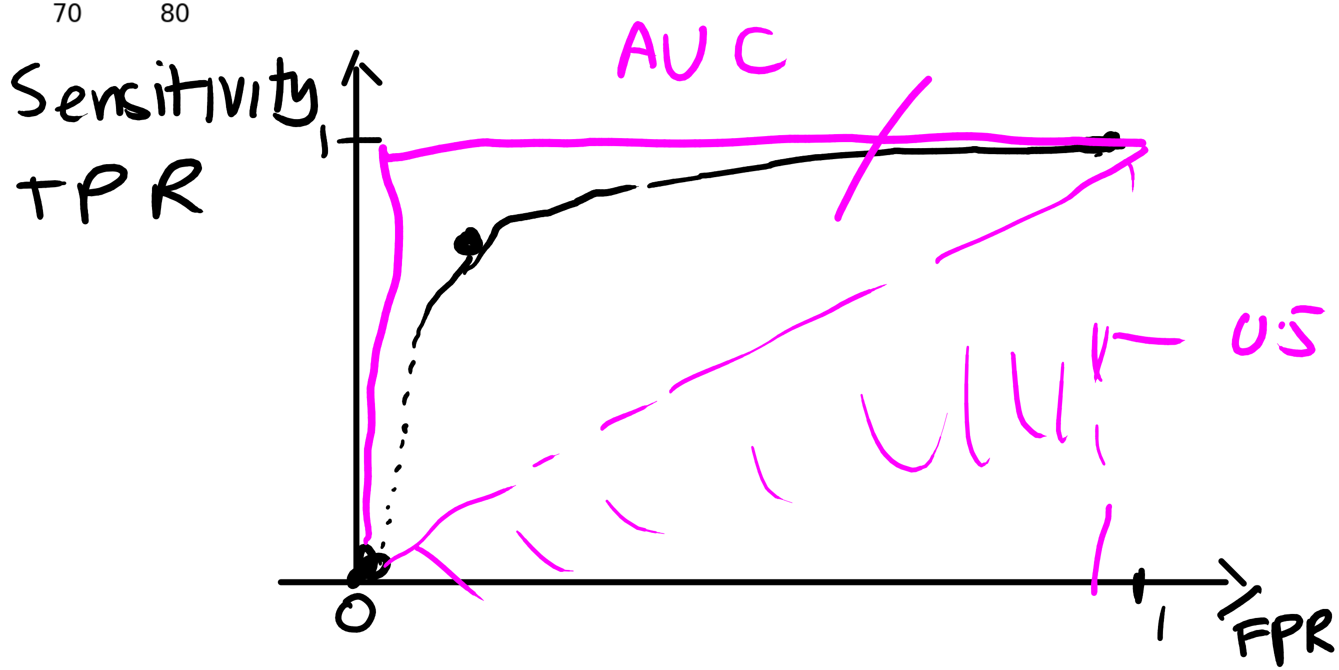
Receiver-operator characteristic



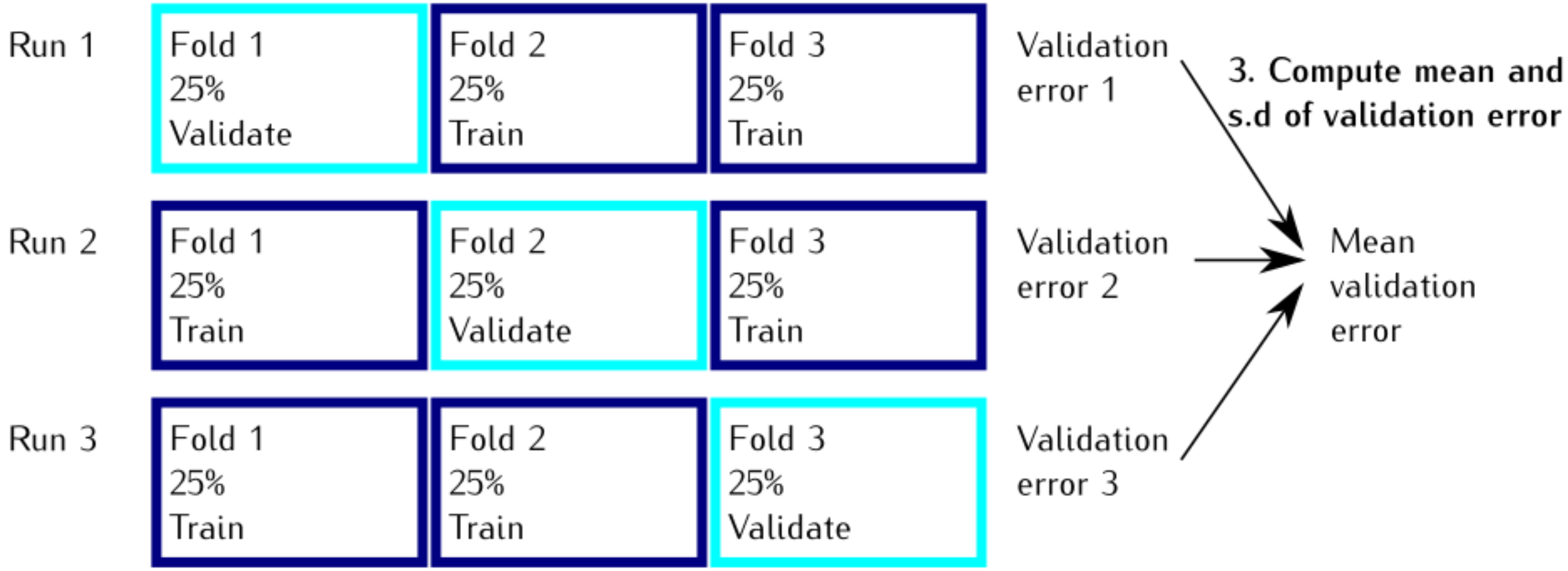
Sensitivity $\dfrac{TP}{TP + FN}$

Selectivity/
Specificity $\dfrac{TN}{TN + FP}$

False positive rate $1 - \dfrac{TN}{TN + FP}$



Cross validation for predicting metrics



c.f. Chapter 12 of the lecture notes

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

- Practiccal and ethical pros and cons of logistic regression versus other methods