Classification and Logistic Regression

CS556

Linear classifiers

Linear classifiers

Linear classifiers

Linear classifiers

Linear classifiers

Framingham health study

Linear classifiers

- Huge impact!

- Used everywhere: marketing, online industry, health care…

- Notable examples: spam filters, all sorts of online services, Framingham health study, assessing health care

Linear classifiers in a nutshell

Hamburger was awesome … service was awful… price was awesome

Pizza was awesome … hot dog was awful…price was awesome

Pizza was awful … hot dog was awesome…price was awesome

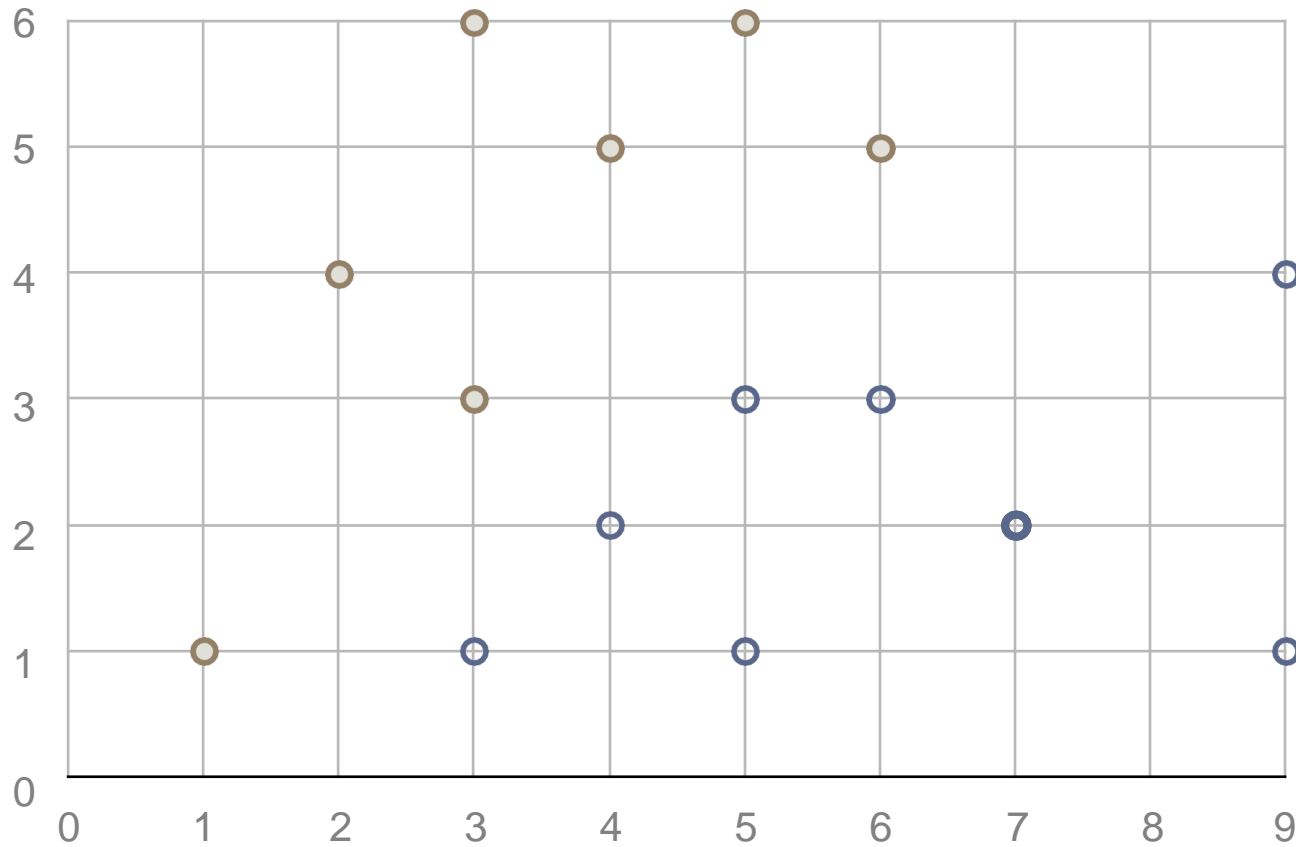Pizza was awful … hot dog was awful…price was awful

Sentiment analysis

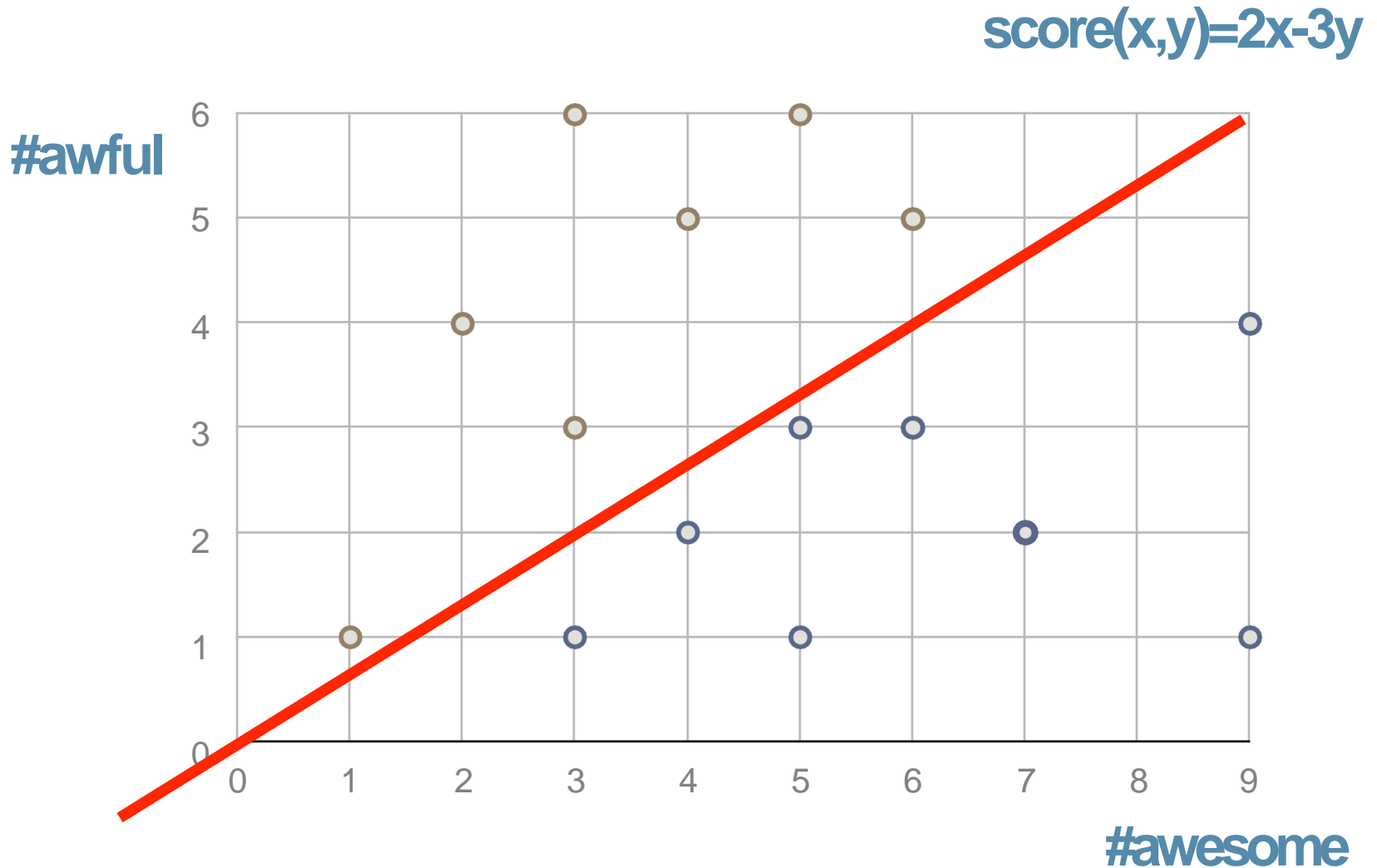|	#awful	#awesome	sentiment
pizza was awful… ice cream was awesome…price was awesome	1	2	+
hot dog was awesome…martini was awesome… price was awful	1	2	+
everything was awful	1	0	-
wine was awful..pasta was awful…dessert was awful..view was	3	1	-

Linear classifiers

#awful

#awesome

Linear classifiers

score(x,y)=2x-3y

#awful

#awesome

14

Linear classifiers

score(x,y)=2x-3y

#awful

#awesome

-12

-9

-8

1

3

15

15

Linear classifiers

score(x,y)=2x-3y

#awful

#awesome

.005
.16
.2
.39
.51
.63
.88
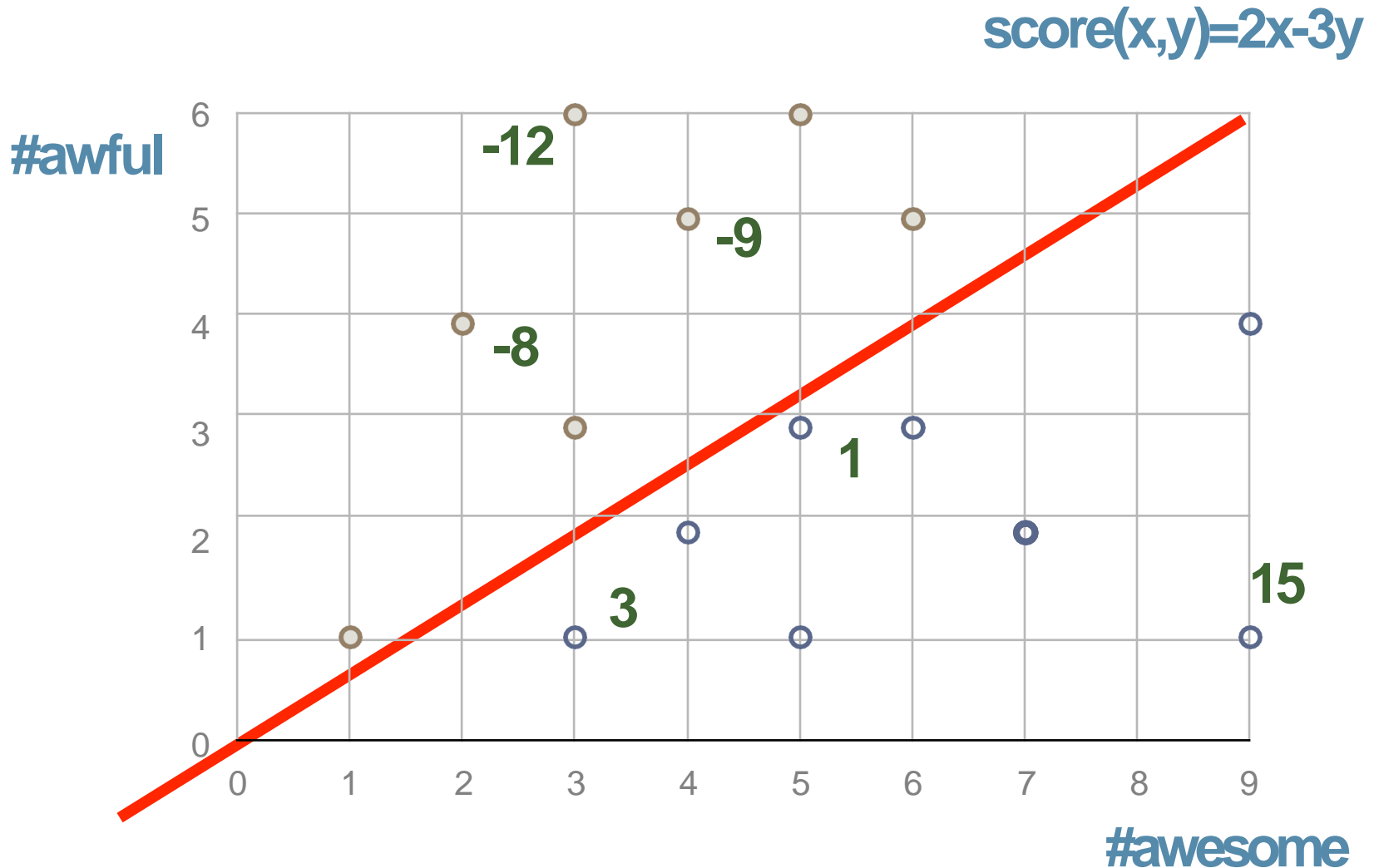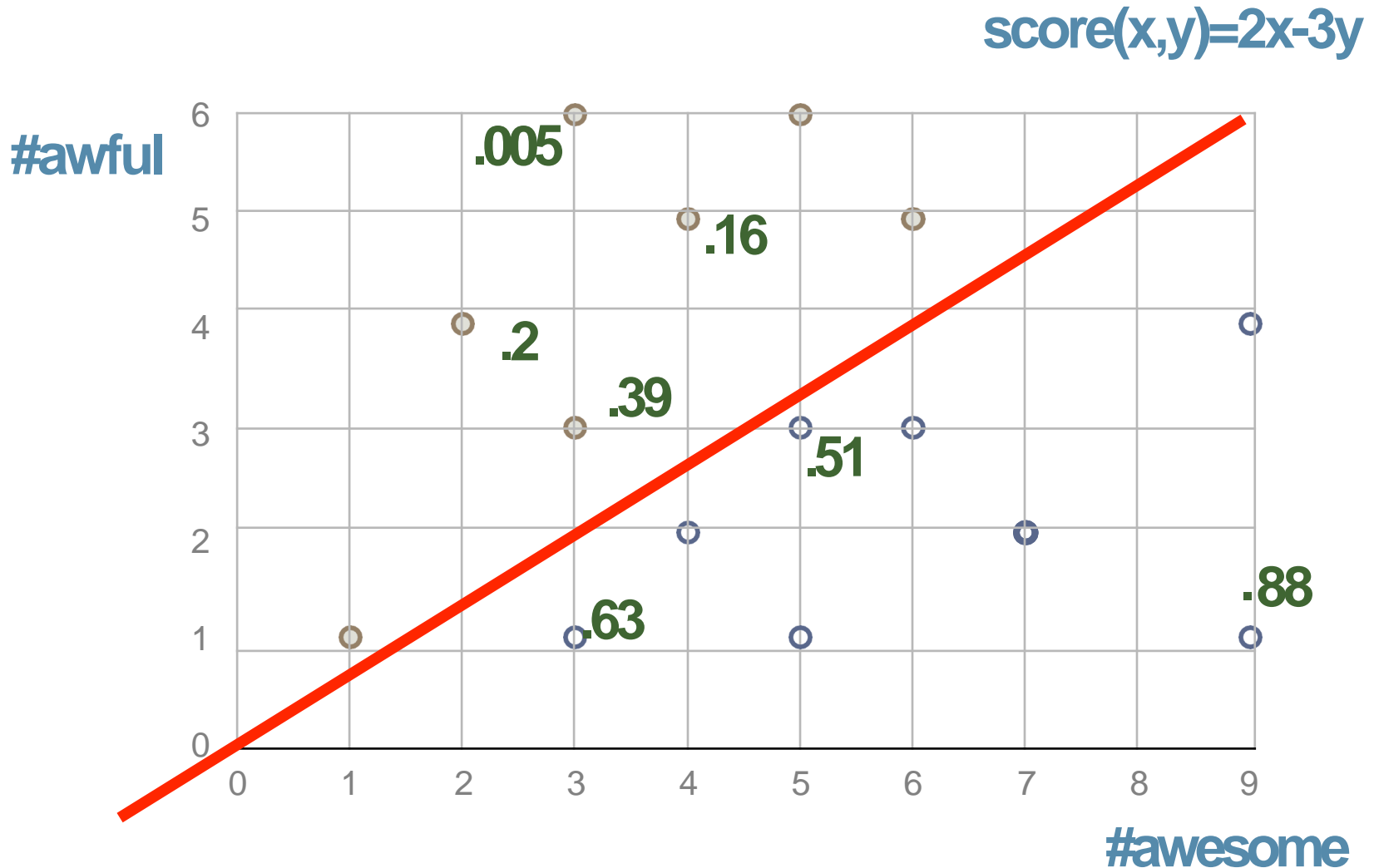
Squeezing scores into probabilities

logistic or sigmoid function
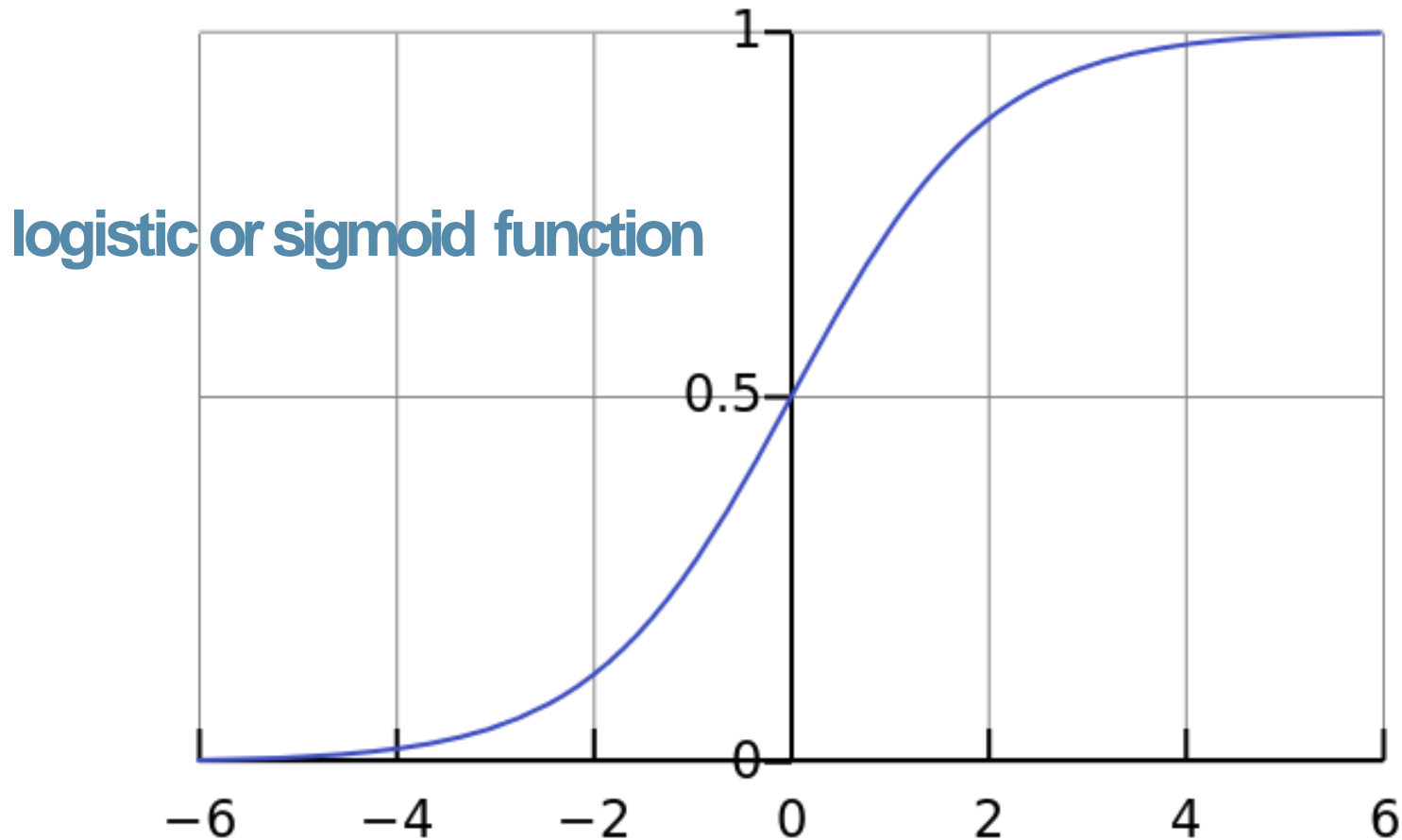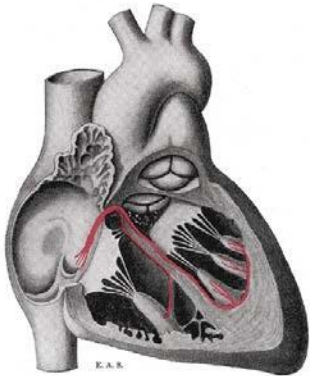
The Framingham Heart Study

Evaluating Risk Factors to Save Lives

Franklin Delano Roosevelt (FDR)

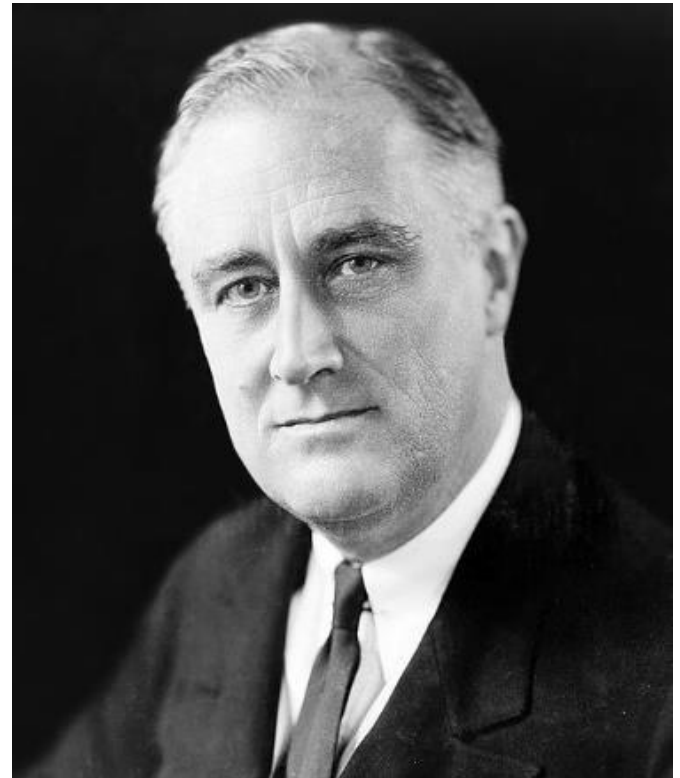- President of the United States, 1933-1945
  - ‣ Longest-serving president
  - ‣ Led country through Great Depression
  - ‣ Commander in Chief of U.S. military in World War II
- Died while president, April 12, 1945

FDR's Blood Pressure

- **Before presidency, blood pressure of 140/100**
  - Healthy blood pressure is less than 120/80
  - Today, this is already considered high blood pressure
- **One year before death, 210/120**
  - Today, this is called Hypertensive Crisis, and emergency care is needed
  - FDR's personal physician:

    "A moderate degree of arteriosclerosis, although no more than normal for a man of his age"
- **Two months before death: 260/150**
- **Day of death: 300/190**

Early Misconceptions

- **High blood pressure dubbed *essential hypertension***

  - Considered important to force blood through arteries

  - Considered harmful to lower blood pressure

- **Today, we know better**

"Today, presidential blood pressure numbers like FDR's would send the country's leading doctors racing down hallways … whisking the nation's leader into the cardiac care unit of Bethesda Naval Hospital."

-- Daniel Levy, Framingham Heart Study Director
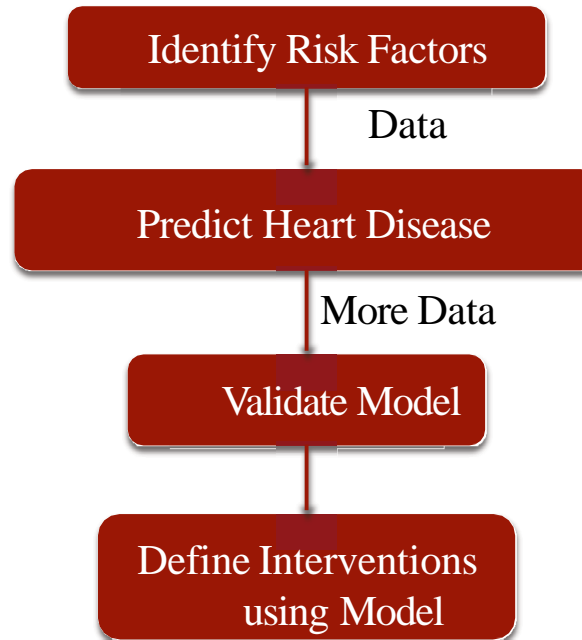
How Did we Learn?

- **In late 1940s, U.S. Government set out to better understand cardiovascular disease (CVD)**

- **Plan: track large cohort of initially healthy patients over time**

- **City of Framingham, MA selected as site for study**

  - Appropriate size

  - Stable population

  - Cooperative doctors and residents

- **1948: beginning of Framingham Heart Study**

The Framingham Heart Study

- **5,209 patients aged 30-59 enrolled**

- **Patients given questionnaire and exam every 2 years**

  - Physical characteristics

  - Behavioral characteristics

  - Test results

- **Exams and questions expanded over time**

- **We will build models using the Framingham data to predict and prevent heart disease**

Analytics to Prevent Heart Disease

Identify Risk Factors

Data

Predict Heart Disease

More Data

Validate Model

Define Interventions using Model

Coronary Heart Disease (CHD)

- **We will predict 10-year risk of CHD**
  - Subject of important 1998 paper, introducing the Framingham Risk Score

- **CHD is a disease of the blood vessels supplying the heart**

- **Heart disease has been the leading cause of death worldwide since 1921**
  - 7.3 million people died from CHD in 2008
  - Since 1950, age-adjusted death rates have declined 60%

Risk Factors

- *Risk factors* are variables that increase the chances of a disease

- **Term coined by William Kannell and Roy Dawber from the Framingham Heart Study**

- **Key to successful prediction of CHD: identifying important risk factors**

Hypothesized CHD Risk Factors

- **We will investigate risk factors collected in the first data collection for the study**

  - Anonymized version of original data

- **Demographic risk factors**

  - *male*: sex of patient

  - *age*: age in years at first examination

  - *education*: Some high school (1), high school/GED (2), some college/vocational school (3), college (4)

Hypothesized CHD Risk Factors

- **Behavioral risk factors**

  - *currentSmoker, cigsPerDay*: Smoking behavior

- **Medical history risk factors**

  - *BPmeds*: On blood pressure medication at time of first examination

  - *prevalentStroke*: Previously had a stroke

  - *prevalentHyp*: Currently hypertensive

  - *diabetes*: Currently has diabetes

Hypothesized CHD Risk Factors

- **Risk factors from first examination**

  - *totChol*: Total cholesterol (mg/dL)

  - *sysBP*: Systolic blood pressure

  - *diaBP*: Diastolic blood pressure

  - *BMI*: Body Mass Index, weight (kg)/height (m)^2

  - *heartRate*: Heart rate (beats/minute)

  - *glucose*: Blood glucose level (mg/dL)

An Analytical Approach

- **Randomly split patients into training and testing sets**

- **Use logistic regression on training set to predict whether or not a patient experienced CHD within 10 years of first examination**

- **Evaluate predictive power on test set**

Logistic Regression

Predicting 10-year CHD

- The independent variables are the risk factors

- The dependent variable is modeled as a binary variable
  - 1 if CHD in 10 years, 0 if not

- This is a *categorical variable*

  - Two (or few) possible outcomes

Logistic Regression

- Predicts the probability of 10-year CHD

  - Denote dependent variable "10-year CHD" by y

  - $P(y = 1)$

- Then $P(y = 0) = 1 - P(y = 1)$

- Independent variables (risk factors): $x_1, x_2, \ldots, x_k$

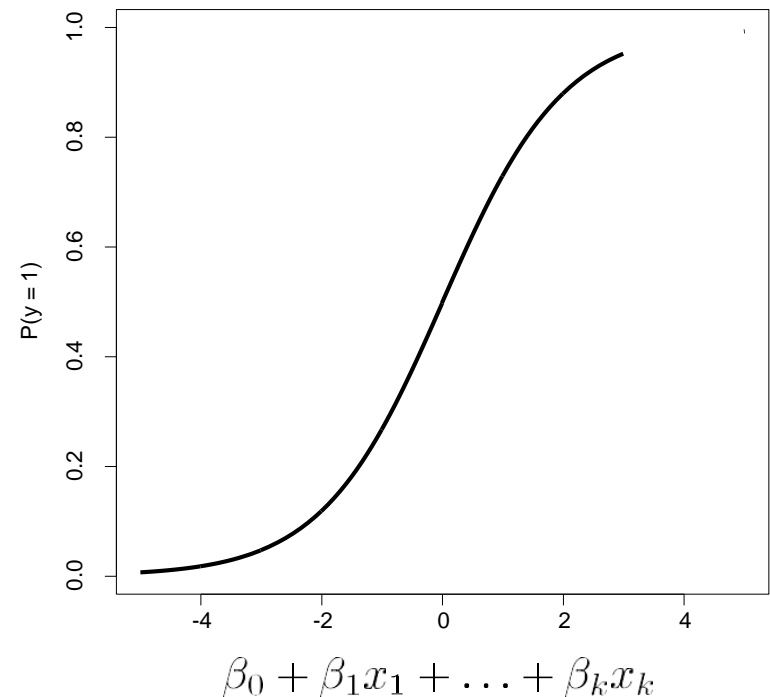- Uses the Logistic Response Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}$$

  - Nonlinear transformation of linear regression equation to produce number between 0 and 1

Understanding the Logistic Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}$$

- Positive values are predictive of class 1

- Negative values are predictive of class 0

$$\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Understanding the Logistic Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}$$

- The coefficients are selected to
  - Predict a high probability for the cases of 10-year CHD
  - Predict a low probability for the cases of no 10-year CHD

Understanding the Logistic Function

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}}$$

- We can instead talk about Odds (like in gambling)

$$\text{Odds} = \frac{P(y = 1)}{P(y = 0)}$$

- Odds > 1 if y = 1 is more likely
- Odds < 1 if y = 0 is more likely
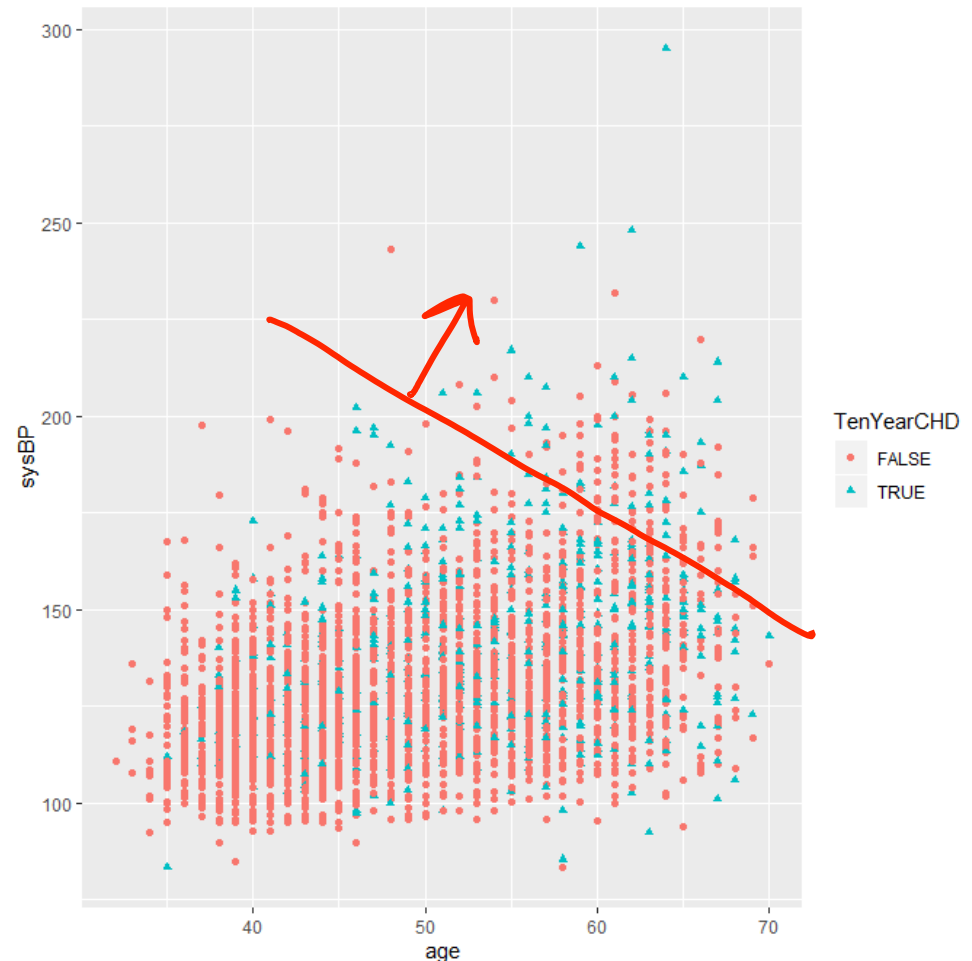
The Logit

- It turns out that

$$\text{Odds} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}$$

$$log(\text{Odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- This is called the "Logit" and looks like linear regression

- The bigger the Logit is, the bigger $P(y = 1)$

Model for Healthcare Quality

- Plot of the independent variables
  - age
  - sysBP
- Red are no 10-year CHD
- Cyan are cases of 10-year CHD

Threshold Value

- The outcome of a logistic regression model is a probability

- Often, we want to make a binary prediction
  - Will the patient incur in 10-year CHD?

- We can do this using a *threshold value* t

- If P(10-year CHD = 1) ≥ t, predict poor quality
- If P(10-year CHD = 1) < t, predict good quality

- What value should we pick for t?

Threshold Value

- Often selected based on which errors are "better"

- If t is large, predict 10-year CHD rarely (when P(y=1) is large)

  - More errors where we say no risk, but there may be risk of 10-year CHD

  - Detects "accurately" patients who would most likely incur into 10-year CHD

- If t is small, predict 10-year CHD commonly (also for P(y=1) small)

  - More errors where we say 10-year CHD, but there may be no risk

  - Detects all patients who might be incurring in 10-year CHD (high recall)

- With no preference between the errors, select t = 0.5

  - Predicts the more likely outcome

Evaluation of classifiers

Selecting a Threshold Value

- Compare actual outcomes to predicted outcomes using a *confusion matrix (classification matrix)*

	Predicted = 0	**Predicted = 1**
Actual = 0	True Negatives (TN)	False Positives (FP)
Actual = 1	False Negatives (FN)	True Positives (TP)

N = number of observations

Overall accuracy = $(TN + TP)/N$ Overall error rate = $(FP + FN)/N$

Sensitivity = $TP/(TP + FN)$ = True Positive Rate

Specificity = $TN/(TN + FP)$ = $1 -$ False Positive Rate

Model Strength

- Model rarely predicts 10-year CHD risk above 50%

    - Accuracy very near a baseline of always predicting no CHD

- Model can differentiate low-risk from high-risk patients (AUC = 0.74)

- Some significant variables suggest interventions

    - Smoking

    - Cholesterol

    - Systolic blood pressure

    - Glucose

Risk Model Validation

- So far, we have used *internal validation*

  - Train with some patients, test with others

- Weakness: unclear if model generalizes to other populations

- Framingham cohort white, middle class

- Important to test on other populations
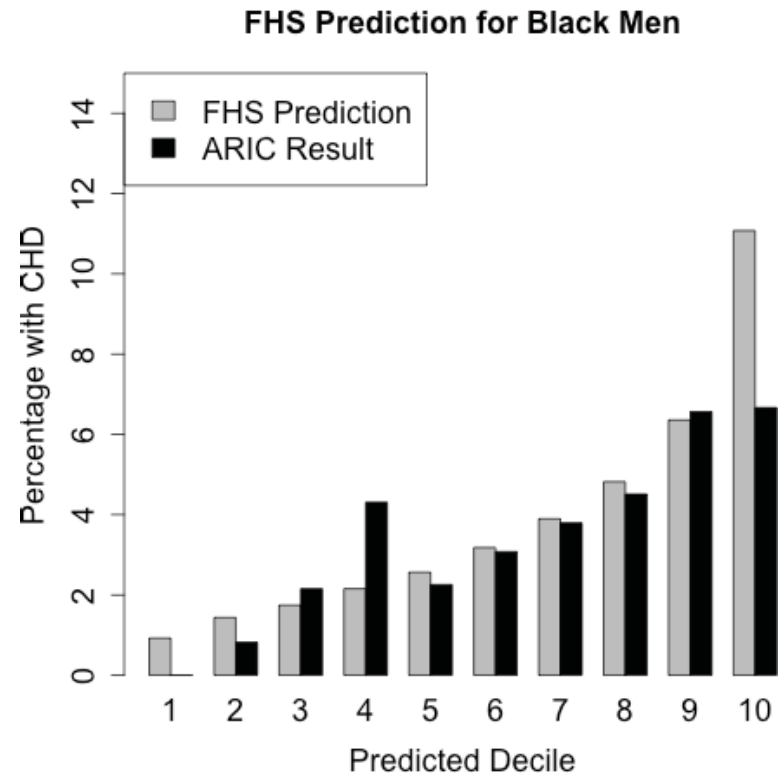
Framingham Risk Model Validation

- Framingham Risk Model tested on diverse cohorts

Study	Population
Atherosclerosis Risk in Communities (ARIC) Study	White and Black
Honolulu Heart Program (HHP)	Japanese American
Puerto Rico Heart Health Program (PR)	Hispanic
Strong Heart Study (SHS)	Native American

- Cohort studies collecting same risk factors

- Validation Plan

  - Predict CHD risk for each patient using FHS model

 - Compare to actual outcomes for each risk decile

Validation for Black Men

- 1,428 black men in ARIC study

- Similar clinical characteristics, except higher diabetes rate

- Similar CHD rate

- Framingham risk model predictions accurate

FHS Prediction for Black Men

Validation for Japanese American Men

- 2,755 Japanese American men in HHS

- Lower CHD rate

- Framingham risk model systematically overpredicts CHD risk

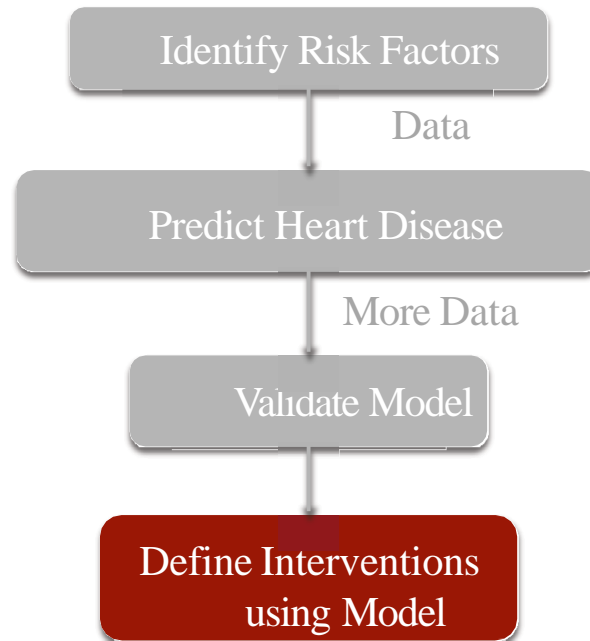FHS Prediction for Japanese American Men
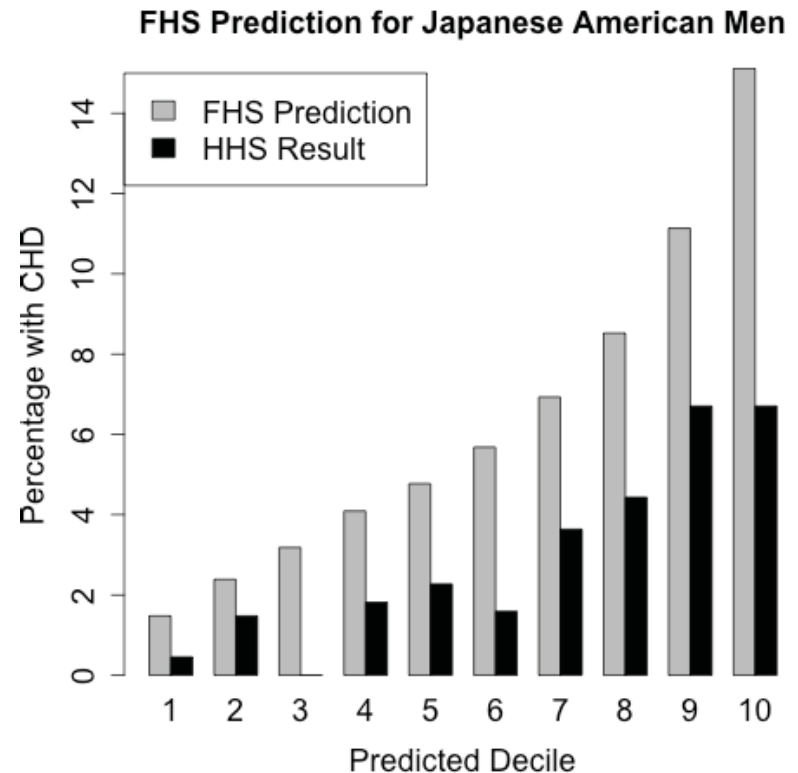
Recalibrated Model

- Recalibration adjusts model to new population

- Changes predicted risk, but does not reorder predictions

- More accurate risk estimates

Adj. FHS Prediction for Japanese American Men
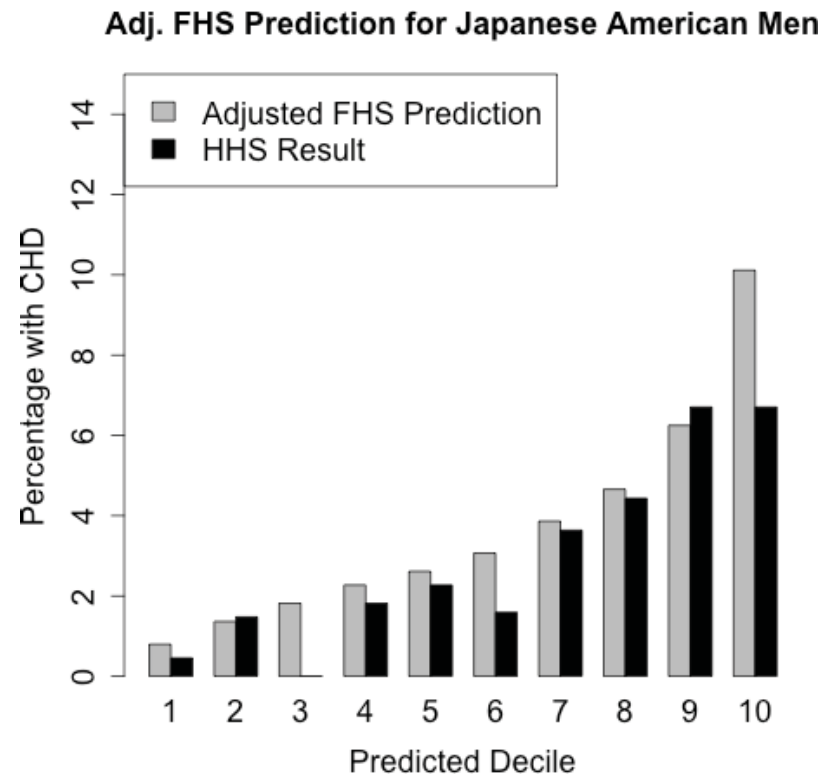
Interventions

Drugs to Lower Blood Pressure

- In FDR's time, hypertension drugs too toxic for practical use

- In 1950s, the diuretic chlorothiazide was developed

- Framingham Heart Study gave Ed Freis the evidence needed to argue for testing effects of BP drugs

- Veterans Administration (VA) Trial: randomized, double blind clinical trial

- Found decreased risk of CHD

- Now, >$1B market for diuretics worldwide

Drugs to Lower Cholesterol

- Despite Framingham results, early cholesterol drugs too toxic for practical use

- In 1970s, first statins were developed

- Study of 4,444 patients with CHD: statins cause 37% risk reduction of   second heart attack

- Study of 6,595 men with high cholesterol: statins cause 32% risk reduction of CVD deaths

- Now, > $20B market for statins worldwide

The Heart Study Through the Years

- More than 2,400 studies use Framingham data

- Many other risk factors evaluated

  - Obesity

  - Exercise

  - Psychosocial issues

  - …

- *Texas Heart Institute Journal*: top 10 cardiology advances of 1900s

Framingham Heart Study Publications by Year
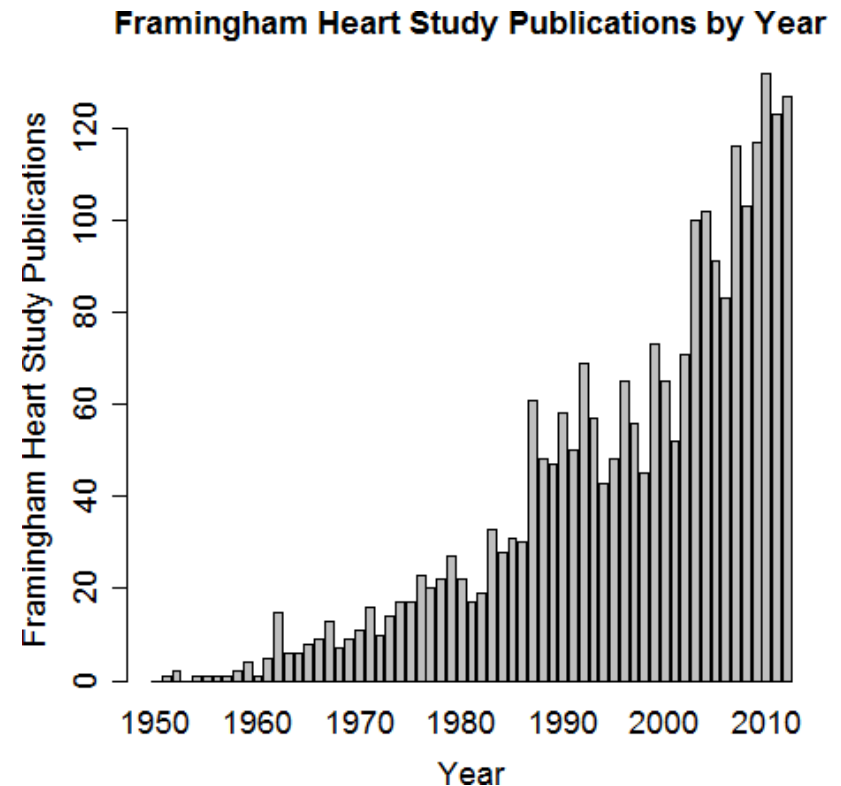
Available Online

Risk Assessment Tool for Estimating Your 10-year Risk of Having a Heart Attack

The risk assessment tool below uses information from the Framingham Heart Study to predict a person's chance of having a heart attack in the next 10 years. This tool is designed for adults aged 20 and older who do not have heart disease or diabetes. To find your risk score, enter your information in the calculator below.

Age: ☐ years
Gender: ○ Female ○ Male
Total Cholesterol: ☐ mg/dL
HDL Cholesterol: ☐ mg/dL
Smoker: ○ No ○ Yes
Systolic Blood Pressure: ☐ mm/Hg
Are you currently on any medication to treat high blood pressure. ○ No ○ Yes

[Calculate Your 10-Year Risk]

↑TOP **Total cholesterol** - Total cholesterol is the sum of all the cholesterol in your blood. The higher your total cholesterol, the greater your risk for heart disease. Here are the total values that matter to you:

Less than 200 mg/dL 'Desirable' level that puts you at lower risk for heart disease. A cholesterol level of 200 mg/dL or greater increases your risk.

200 to 239 mg/dL 'Borderline-high.'

23

Clinical Decision Rules

- Paved the way for *clinical decision rules*

- Predict clinical outcomes with data

  - Patient and disease characteristics

  - Test results

- More than 75,000 published across medicine

- Rate increasing

Prediction Rules Published Each Year
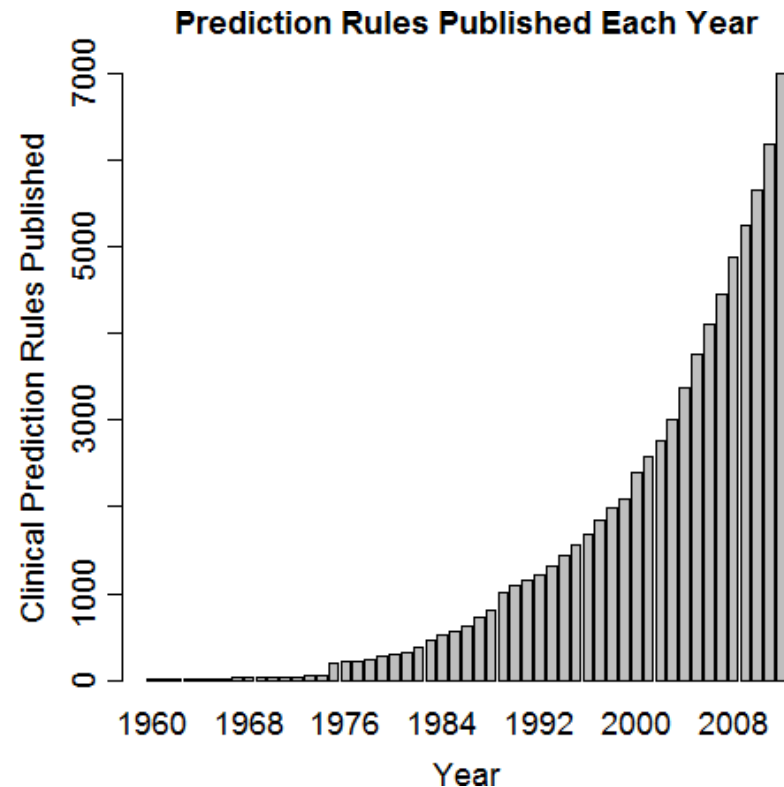
Notes on Learning Logistic Regression

Maximum Likelihood Estimation

- We use maximum likelihood estimation (MLE) to estimate the parameters of the logistic regression

- The labels that we are predicting are binary, and the output of our logistic regression function is supposed to be the probability that the label is one

- This means that we can (and should) interpret each label as a Bernoulli random variable: Y ~ Ber(p) where $p = \sigma(\theta^T x)$

- We write then the probability of one data point as:

$$P(Y = y | X = \mathbf{x}) = \sigma(\theta^T \mathbf{x})^y \cdot \left[1 - \sigma(\theta^T \mathbf{x}) \right]^{(1-y)}$$

Maximum Likelihood Estimation

- So the likelihood of all data is:

$$L(\theta) = \prod_{i=1}^{n} P(Y = y^{(i)} \mid X = \mathbf{x}^{(i)})$$

$$= \prod_{i=1}^{n} \sigma(\theta^T \mathbf{x}^{(i)})^{y^{(i)}} \cdot \left[1 - \sigma(\theta^T \mathbf{x}^{(i)})\right]^{(1-y^{(i)})}$$

- Taking the log, we get the log likelihood for logistic regression

$$LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \sigma(\theta^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log[1 - \sigma(\theta^T \mathbf{x}^{(i)})]$$

- to maximize with numerical techniques such as

  ‣ Line Search

  ‣ Simulated Annealing

  ‣ Gradient Descent

  ‣ Newton's Method

Regularization

- It would still be appropriate to regularize the optimization with an L2 (or L1) loss on the parameter values, to penalize very large coefficients and avoid overfitting:

$$cost = -\sum_{i=1}^{n}\{y^{(i)} \log \sigma(\theta^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log[1 - \sigma(\theta^T \mathbf{x}^{(i)})]\} + \lambda \sum \theta^2$$

Thank you

Slide credits:
- MIT open course ware
- Prof. Alessandro Panconesi
- Will Monroe's lecture notes at Stanford
- Prof. Fabio Galasso