The Framingham Heart Study

Franklin Delano Roosevelt (FDR)

* President of the United States, 1933 - 1945
  + Longest-serving president
  + Led country through Great Depression
  + Commander in Chief of US 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, presidential blood pressure numbers like FDR’s would send the country’s leading doctors racing down hallways … whisking the nation’s leader into 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 health 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 exams 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

Diagram

Description automatically generated

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 leading the 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 Kannel and Roy Dawber from the Framing Ham 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/vocation school (3), college(4)

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

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 variable 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 color white, middle class
* Important to test on other populations

Framingham Risk Model Validation

* Framingham Risk Model tested on diverse cohorts

Text

Description automatically generated with low confidence

* Cohort studies collecting same risk factors
* Validation Plan
  + Predict CHD risk for each patient using FHS model
  + Compare to actual outcomes for each docile

Drugs to Lower Blood Pressure

* In FDR’s time, hypertension drugs too toxic for practical use
* In 1950s, the diuretic chlorothiazied 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: status 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

Research Directions and Challenges

* Second generation enrolled in 1971, third in 2002
  + Enables study of family history as a risk factor
* More diverse cohorts begun in 1994 and 2003
* Social network analysis of participants
* Genome-wide association study linking studying genetics as risk factors
* Many challenges related to funding
  + Funding cuts in 1969 nearly closed study
  + 2013 sequester threatening to close study

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

Framingham Heart Study in R

Load in the dataset

*# Read in the dataset*

framingham = **read.csv**("framingham.csv")

Examine structure

*# Look at structure*

**str**(framingham)

## 'data.frame': 4240 obs. of 16 variables:

## $ male : int 1 0 1 0 0 0 0 0 1 1 ...

## $ age : int 39 46 48 61 46 43 63 45 52 43 ...

## $ education : int 4 2 1 3 3 2 1 2 1 1 ...

## $ currentSmoker : int 0 0 1 1 1 0 0 1 0 1 ...

## $ cigsPerDay : int 0 0 20 30 23 0 0 20 0 30 ...

## $ BPMeds : int 0 0 0 0 0 0 0 0 0 0 ...

## $ prevalentStroke: int 0 0 0 0 0 0 0 0 0 0 ...

## $ prevalentHyp : int 0 0 0 1 0 1 0 0 1 1 ...

## $ diabetes : int 0 0 0 0 0 0 0 0 0 0 ...

## $ totChol : int 195 250 245 225 285 228 205 313 260 225 ...

## $ sysBP : num 106 121 128 150 130 ...

## $ diaBP : num 70 81 80 95 84 110 71 71 89 107 ...

## $ BMI : num 27 28.7 25.3 28.6 23.1 ...

## $ heartRate : int 80 95 75 65 85 77 60 79 76 93 ...

## $ glucose : int 77 76 70 103 85 99 85 78 79 88 ...

## $ TenYearCHD : int 0 0 0 1 0 0 1 0 0 0 ...

Split the dataset

*# Load the library caTools*

**library**(caTools)

*# Randomly split the data into training and testing sets*

**set.seed**(1000)

split = **sample.split**(framingham**$**TenYearCHD, SplitRatio = 0.65)

*# Split up the data using subset*

train = **subset**(framingham, split**==**TRUE)

test = **subset**(framingham, split**==**FALSE)

Logistic Regression

*# Logistic Regression Model*

framinghamLog = **glm**(TenYearCHD **~** ., data = train, family=binomial)

**summary**(framinghamLog)

##

## Call:

## glm(formula = TenYearCHD ~ ., family = binomial, data = train)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -1.8487 -0.6007 -0.4257 -0.2842 2.8369

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -7.886574 0.890729 -8.854 < 2e-16 \*\*\*

## male 0.528457 0.135443 3.902 9.55e-05 \*\*\*

## age 0.062055 0.008343 7.438 1.02e-13 \*\*\*

## education -0.058923 0.062430 -0.944 0.34525

## currentSmoker 0.093240 0.194008 0.481 0.63080

## cigsPerDay 0.015008 0.007826 1.918 0.05514 .

## BPMeds 0.311221 0.287408 1.083 0.27887

## prevalentStroke 1.165794 0.571215 2.041 0.04126 \*

## prevalentHyp 0.315818 0.171765 1.839 0.06596 .

## diabetes -0.421494 0.407990 -1.033 0.30156

## totChol 0.003835 0.001377 2.786 0.00533 \*\*

## sysBP 0.011344 0.004566 2.485 0.01297 \*

## diaBP -0.004740 0.008001 -0.592 0.55353

## BMI 0.010723 0.016157 0.664 0.50689

## heartRate -0.008099 0.005313 -1.524 0.12739

## glucose 0.008935 0.002836 3.150 0.00163 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 2020.7 on 2384 degrees of freedom

## Residual deviance: 1792.3 on 2369 degrees of freedom

## (371 observations deleted due to missingness)

## AIC: 1824.3

##

## Number of Fisher Scoring iterations: 5

Make Predictions

*# Predictions on the test set*

predictTest = **predict**(framinghamLog, type="response", newdata=test)

*# Confusion matrix with threshold of 0.5*

z = **table**(test**$**TenYearCHD, predictTest **>** 0.5)

**kable**(z)

|  | **FALSE** | **TRUE** |
| --- | --- | --- |
| 0 | 1069 | 6 |
| 1 | 187 | 11 |

*# Accuracy*

(1069**+**11)**/**(1069**+**6**+**187**+**11)

## [1] 0.8483896

*# Baseline accuracy*

(1069**+**6)**/**(1069**+**6**+**187**+**11)

## [1] 0.8444619

AUC

*# Test set AUC*

**library**(ROCR)

ROCRpred = **prediction**(predictTest, test**$**TenYearCHD)

**as.numeric**(**performance**(ROCRpred, "auc")**@**y.values)

## [1] 0.7421095