10 Year Risk of Heart Disease

A logistic regression approach to the Framingham Heart Study Data

Contents

01

INTRODUCTION

click to add text

Framingham Heart Study
1948-present

- Heart disease No. 1 cause of death in U.S.
- Identify risk factors for cardiovascular disease

- ~5000 participants from Framingham, Mass
- Exam/ lab tests/ lifestyle questions
- Track major heart events: chest pain, heart attacks, strokes, death…

- Data subset 1956 - 1968

“Can we predict if a person will develop heart disease within 10 years?”

OUR QUESTION

Can we predict if a person will develop heart disease within 10 years?

- Binary response (Yes/No)
- Logistic Regression

$$\ln \frac{p}{1-p} = X'\beta$$

02

EXPLORING DATA ANALYSIS

Identify risk factors

↓

Data cleaning

↓

Predictors/Responses relationships

Potetial Risk Factors

➢Demographics

- Sex (1707 Men, 2282 Women)
- **Age** (Range: **30-70**)

➢Lifestyle

- Smoking Status (Yes or No)
- Cigarettes per day

➢Medical Exam

- Period (1-3)
- Time (days)
- Systolic BP (mm HG)
- Diastolic BP (mmHG)
- Total Cholesterol (mg/dL)
- Body Mass Index (BMI)
- Glucose (mg/dL)
- Heart rate (per min)

➢Medical History

- Diabetes
- Blood pressure medication
- Prev. Angina pectoris
- Prev. Myocardial infarction
- Prev. Stroke
- Prev. Hypertension

Total: 11,627 observations

Dealing with Missing Values and Multicollinearity

➢Design matrix

- PERIOD=1
  - the most observations available
  - sufficient time to measure 10 year risk

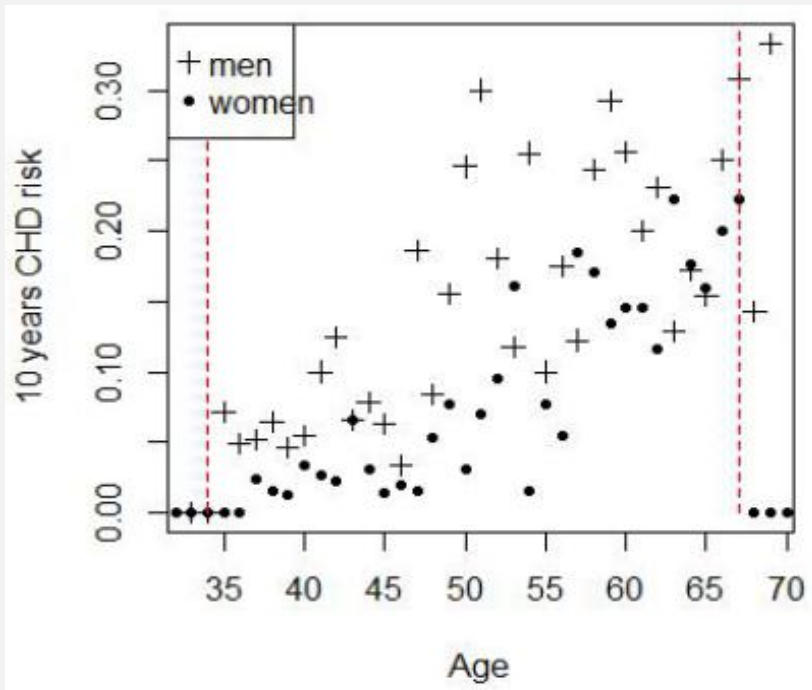
➢Missing Values

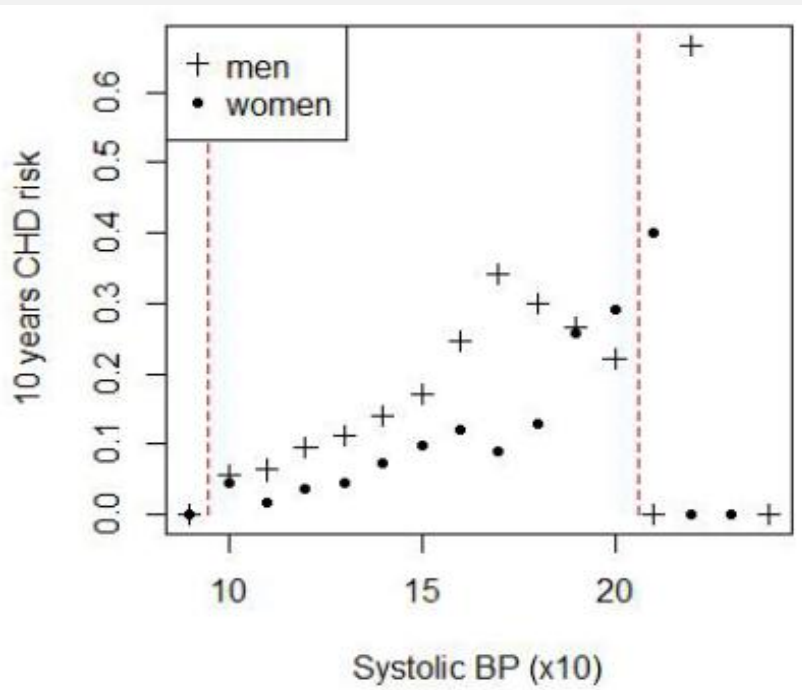
- Remove observation
- Replace value with 0
- Replace value with mean of the column

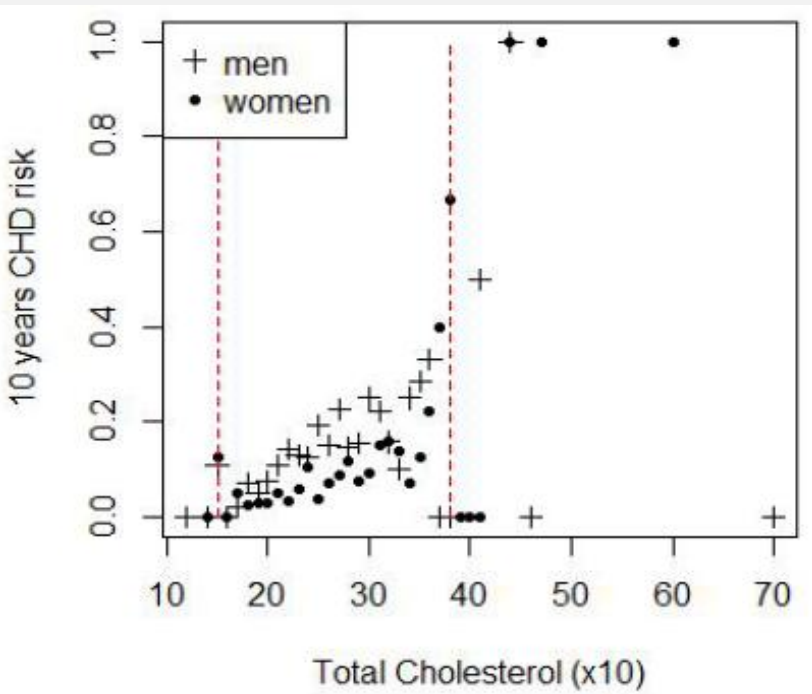
➢Correlation

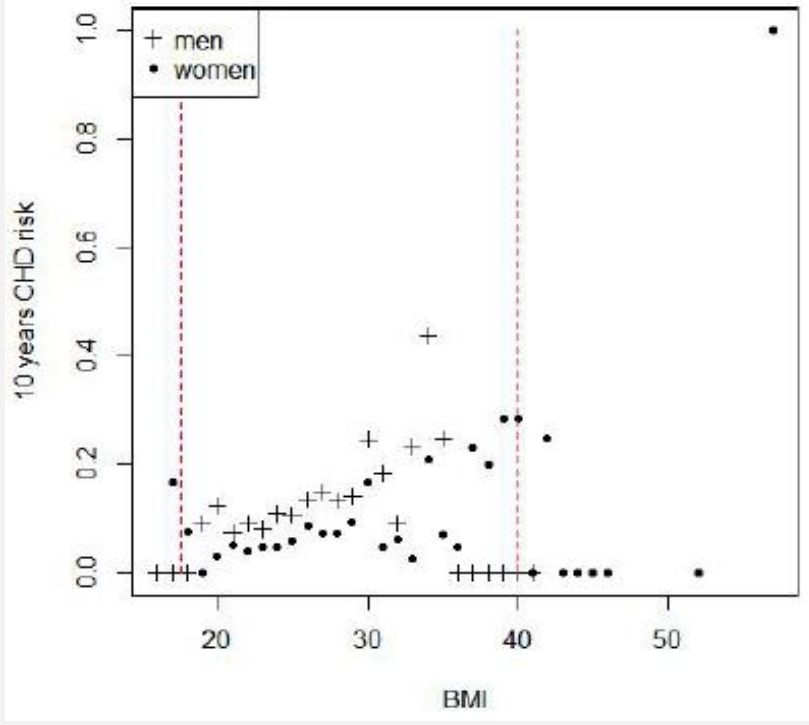
- Systolic BP and diastolic BP (.78)
- Smoking and cigarettes per day (.77)
- Glucose and diabetes (.61)
- Prev hypertension and both systolic and diastolic BP
- (.69 and.62 respectively)

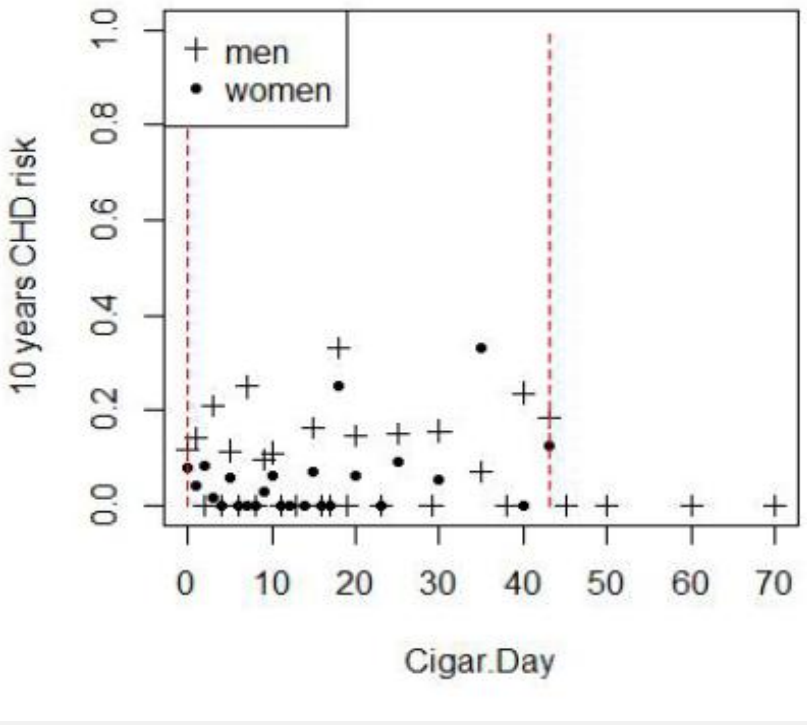
After cleaning: 3,531 observations

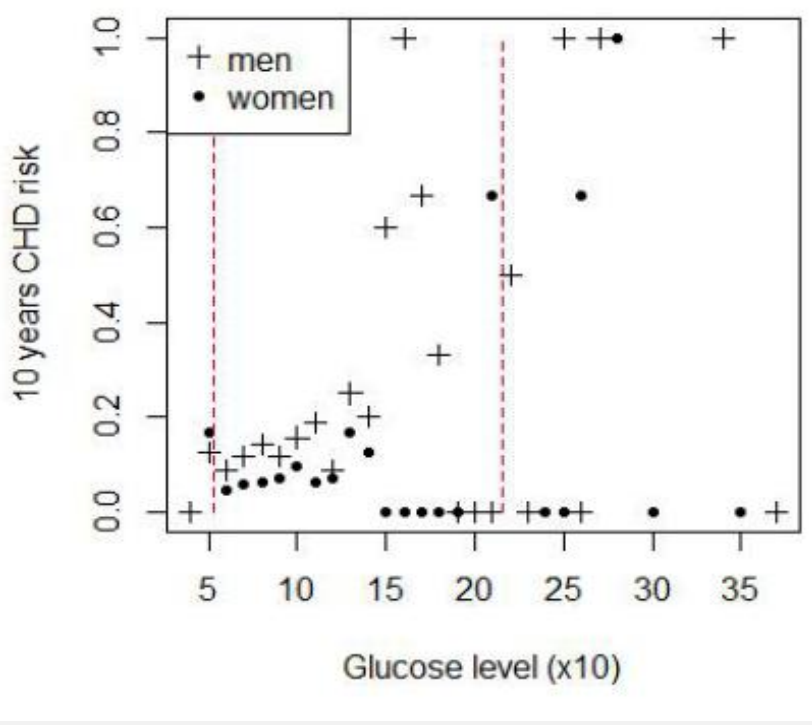

men
women
10 years CHD risk
0.00
0.10
0.20
0.30
35
40
45
50
55
60
65
70
Age

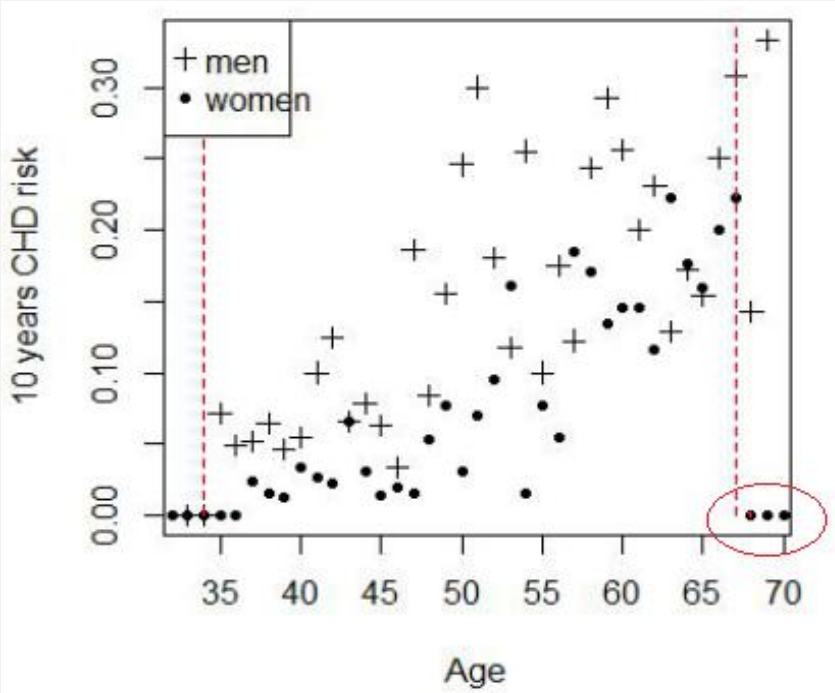

men
women
10 years CHD risk
0.0
0.1
0.2
0.3
0.4
0.5
0.6
10
15
20
Systolic BP (x10)

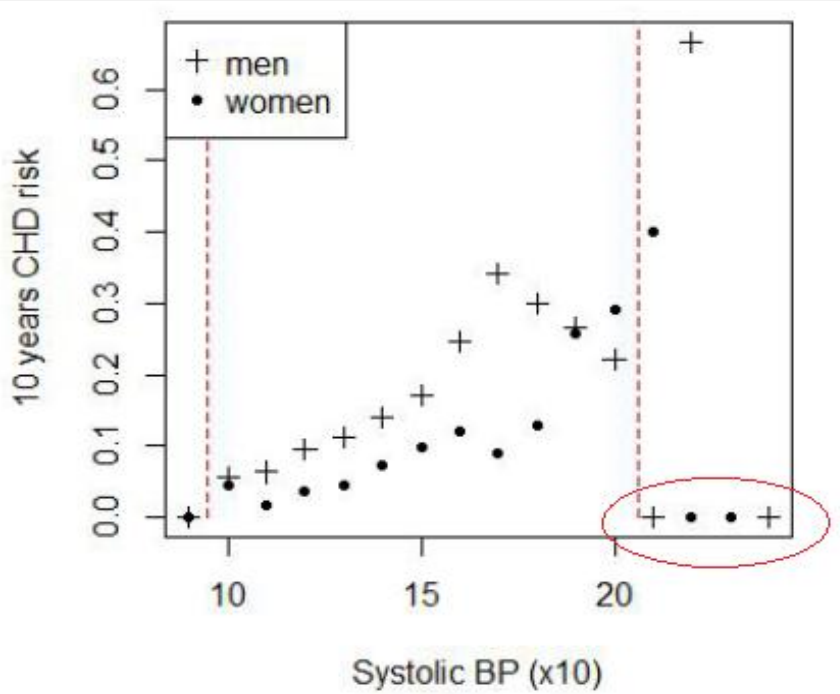

men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
60
70
Total Cholesterol (x10)

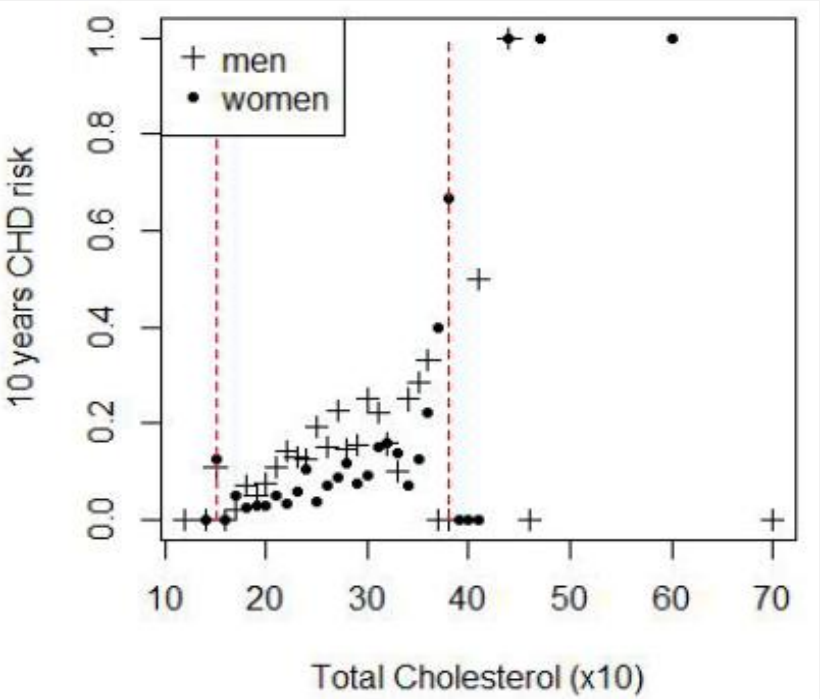

men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
BMI

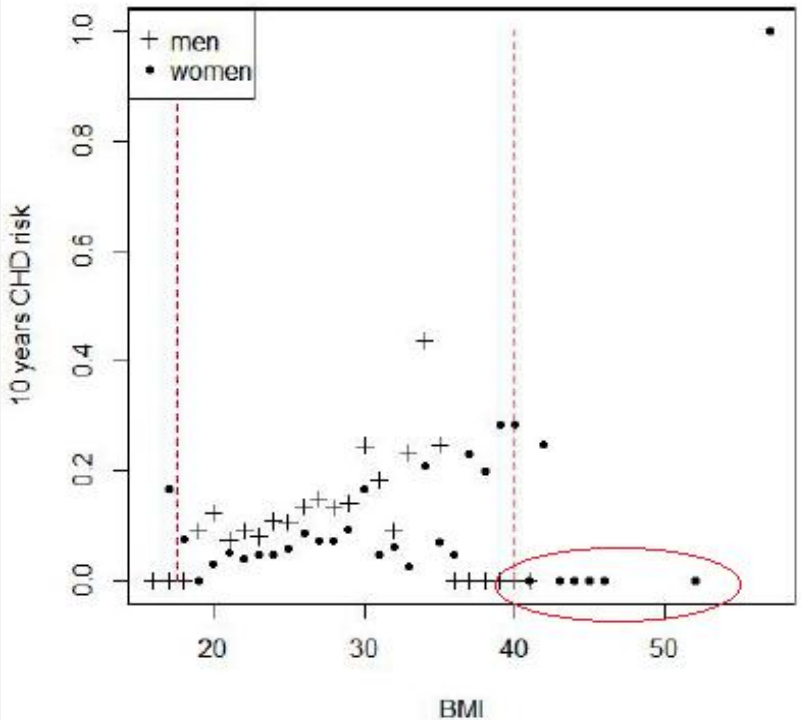

men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Cigar.Day

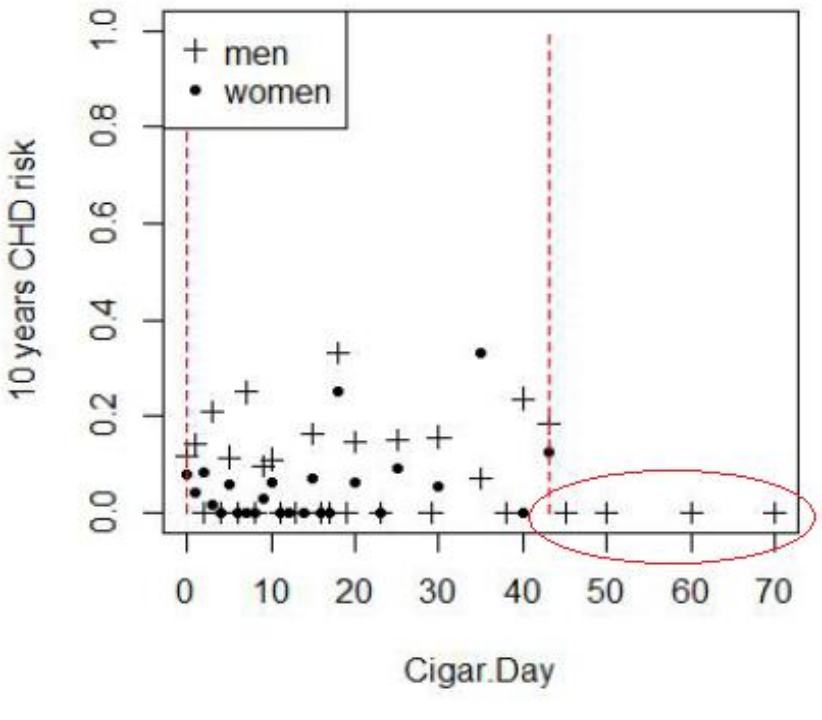

men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
30
35
Glucose level (x10)

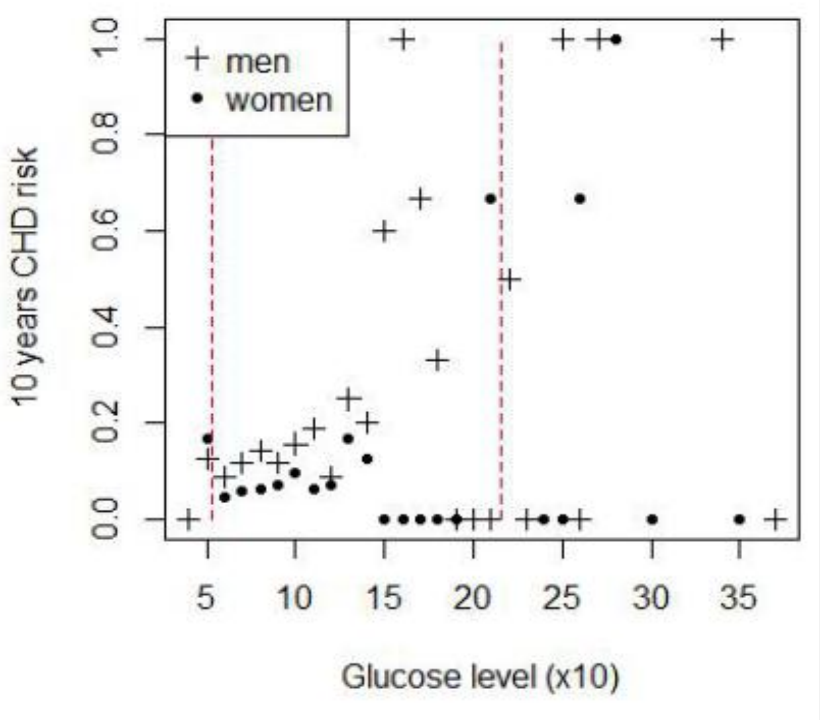

men
women
10 years CHD risk
0.00
0.10
0.20
0.30
35
40
45
50
55
60
65
70
Age

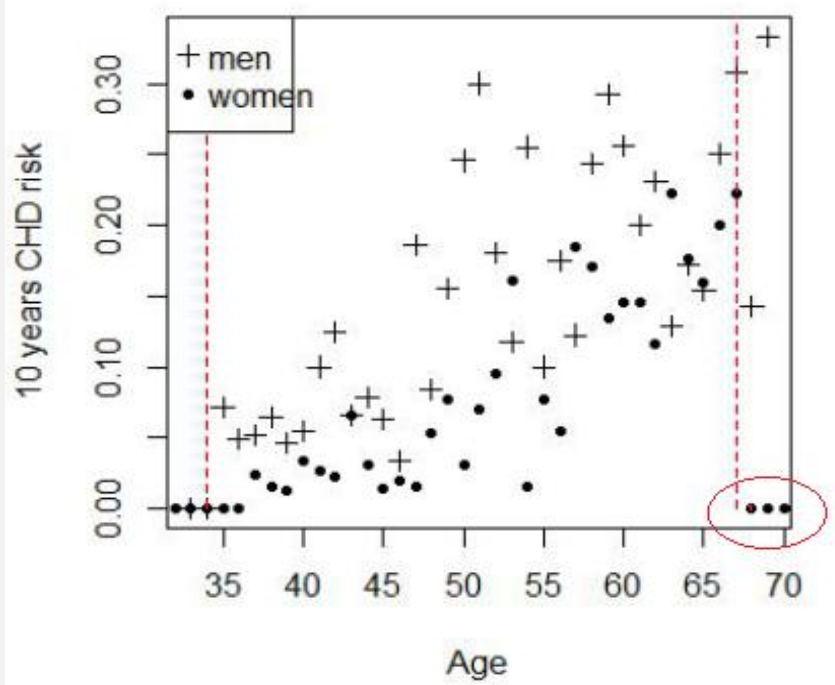

men
women
10 years CHD risk
0.0
0.1
0.2
0.3
0.4
0.5
0.6
10
15
20
Systolic BP (x10)

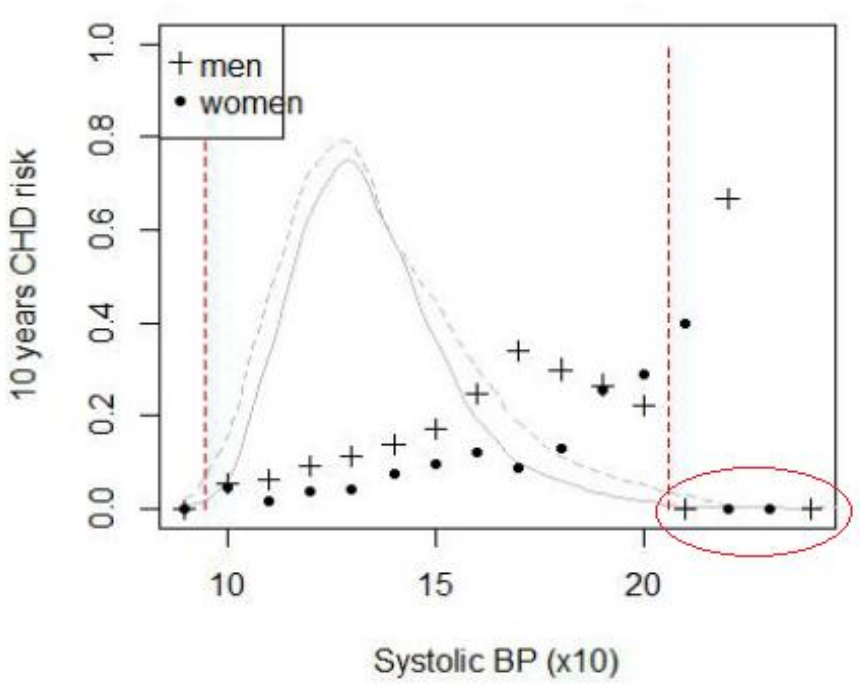

men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
60
70
Total Cholesterol (x10)

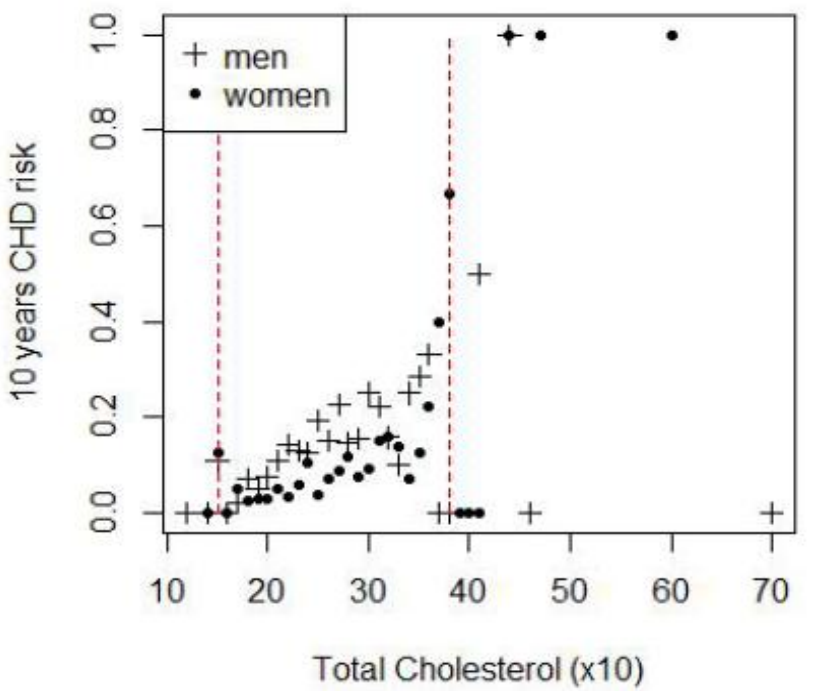

men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
BMI

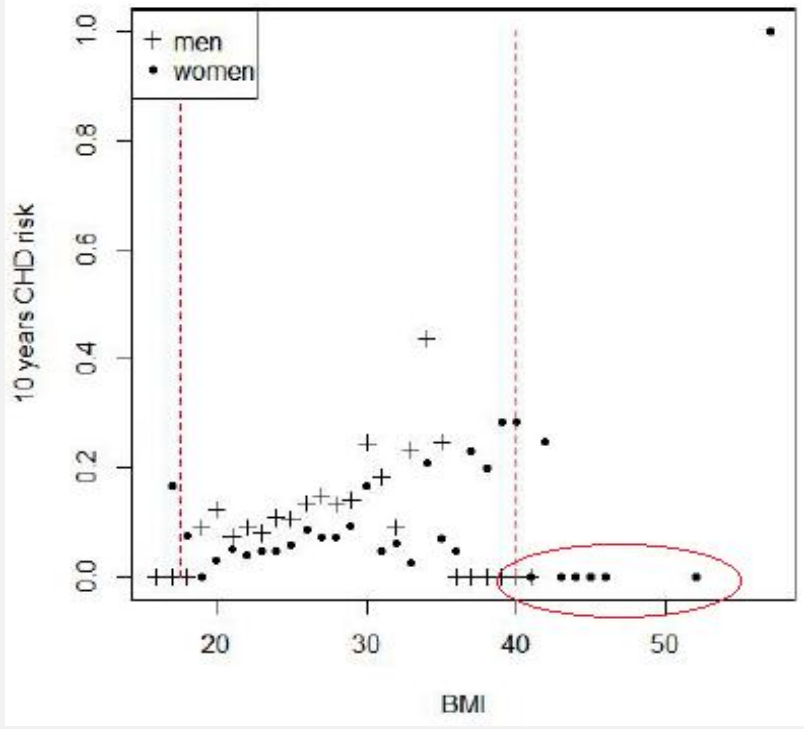

men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Cigar.Day

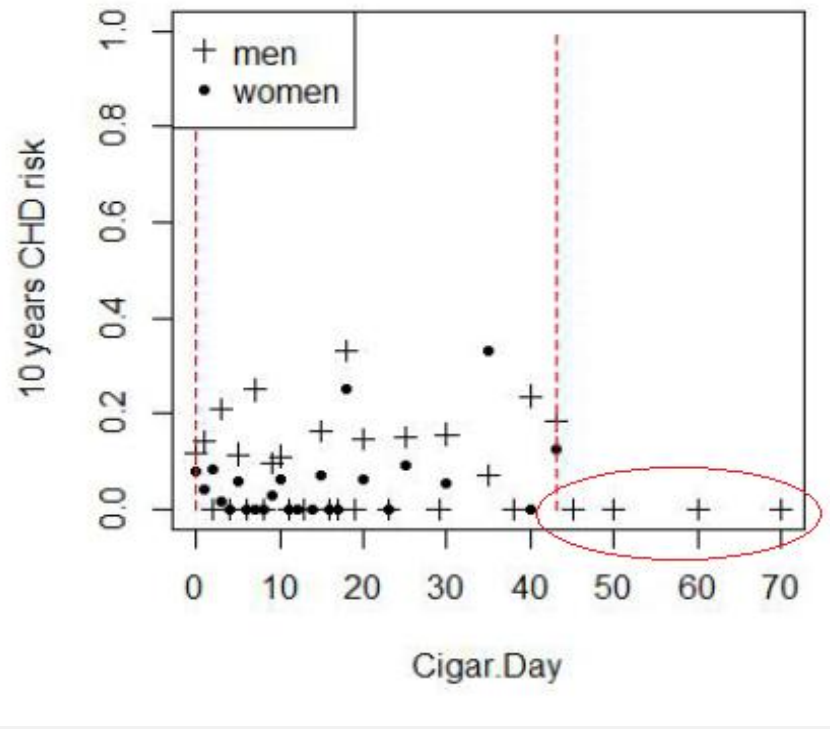

men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
30
35
Glucose level (x10)

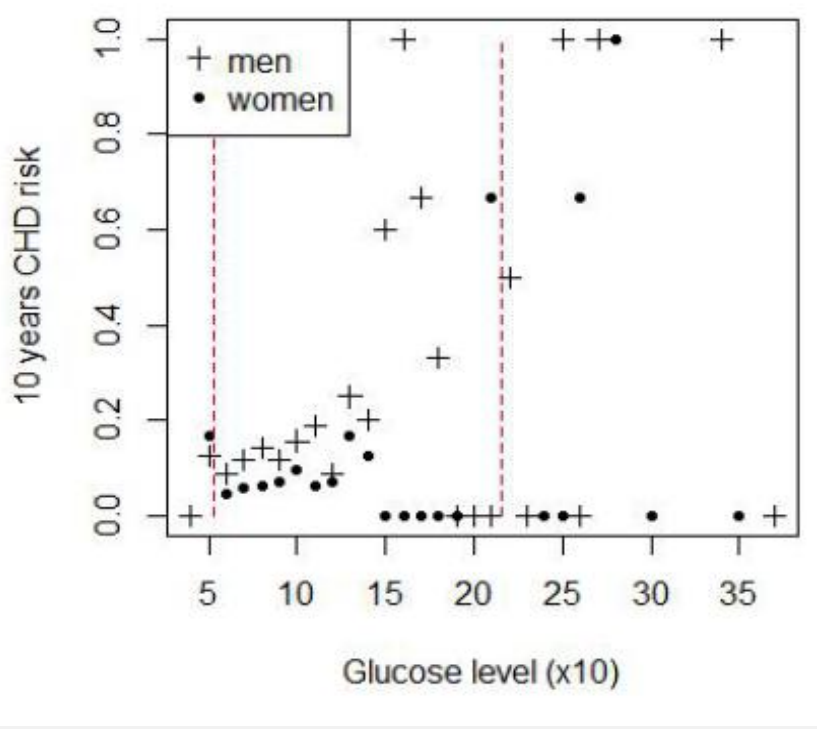

men
women
10 years CHD risk
0.00
0.10
0.20
0.30
35
40
45
50
55
60
65
70
Age


men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
10
15
20
Systolic BP (x10)


men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
60
70
Total Cholesterol (x10)


men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
BMI


men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Cigar.Day


men
women
10 years CHD risk
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
30
35
Glucose level (x10)

03

VARIABLE SELECTION

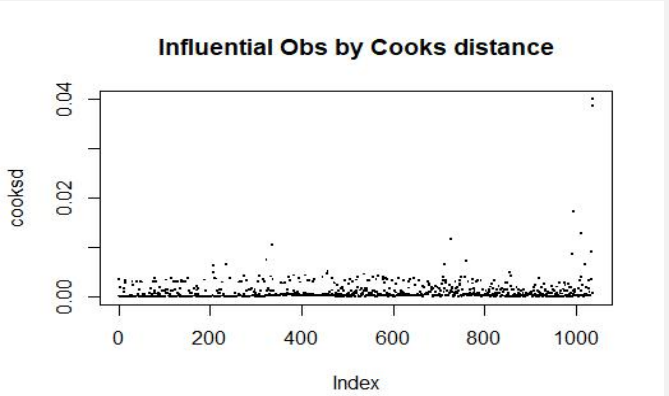
Variable Selection

- Variable transformations, interactions (sex with cig smoked per day,bp and heart rate)
- combining predictors sys Bp and Dia BP.0-normal, (1-elevated, 2-high stage1, 3-high stage2, 4-hypertensive crisis)
- Binning of age groups(30-40,41-50, 51-60 and 6 -70)
- But we found it not significant in getting higher predictive power.
- AIC , Deviance, LRT, and p-value of Pearson Chi Squared test.

Variable selection

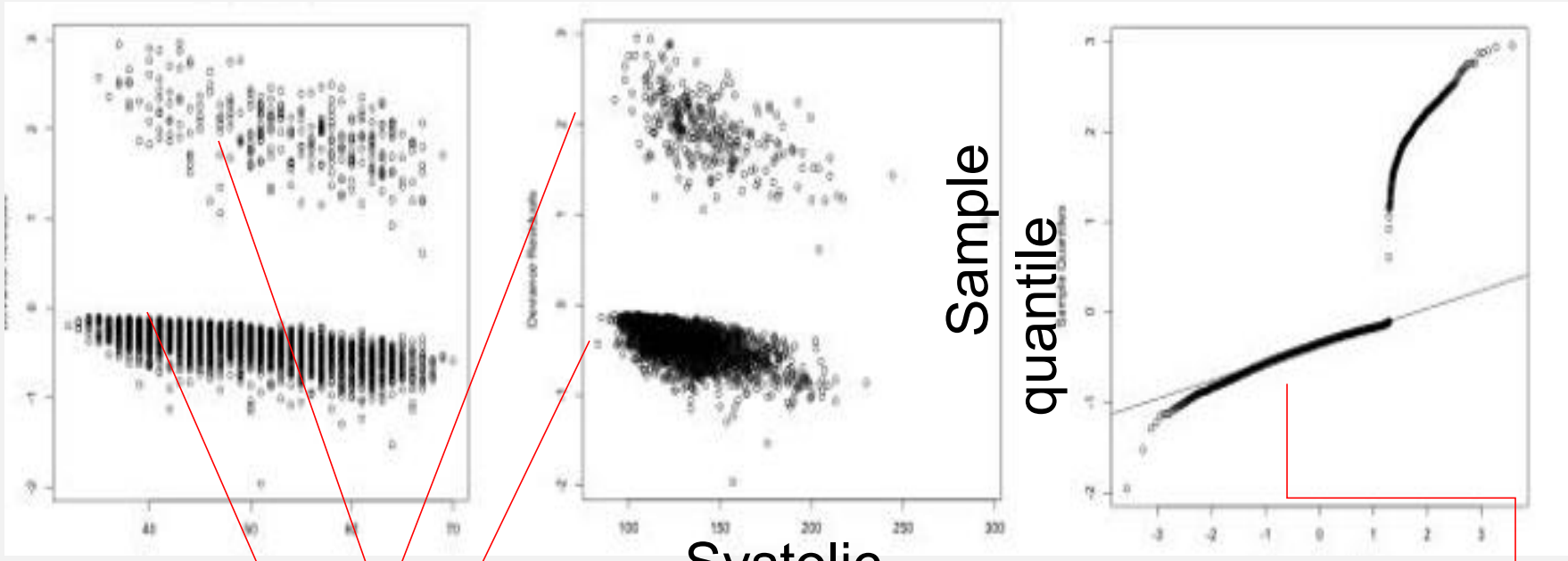
No	k	p	Intercept	HeartRate	BMI	Gluc	Cigar.Day	Tol.Chol	SysBP	Sex	Age	AIC	Deviance	LRT	Pear.Chi
1	8	9	1	1	1	1	1	1	1	1	1	2037.84	2019.84	1.00	0.78
2	7	8	1	0	1	1	1	1	1	1	1	2038.70	2022.70	1.00	0.78
3	6	7	1	0	1	1	1	0	1	1	1	2066.11	2052.11	1.00	0.93
4	7	8	1	1	1	1	1	0	1	1	1	2066.35	2050.35	1.00	0.93

Full Model	Reduce Model	P-value
1	2	0.09104



Event~BMI+Gluc+Cigar.Day+Tol.Chol+SysBP+Sex+Age

Residual plots and qq-plot



Two chunks of data because of binary responses 0 and 1

Normall distributed

04

MODEL VALIDATION

Model Validation

- Generated 80%:20% training to validation split with 2824 and 707 records respectively
- Misclassification error rate: 9.4% on the validation sample
- Model Concordance: 75.62%

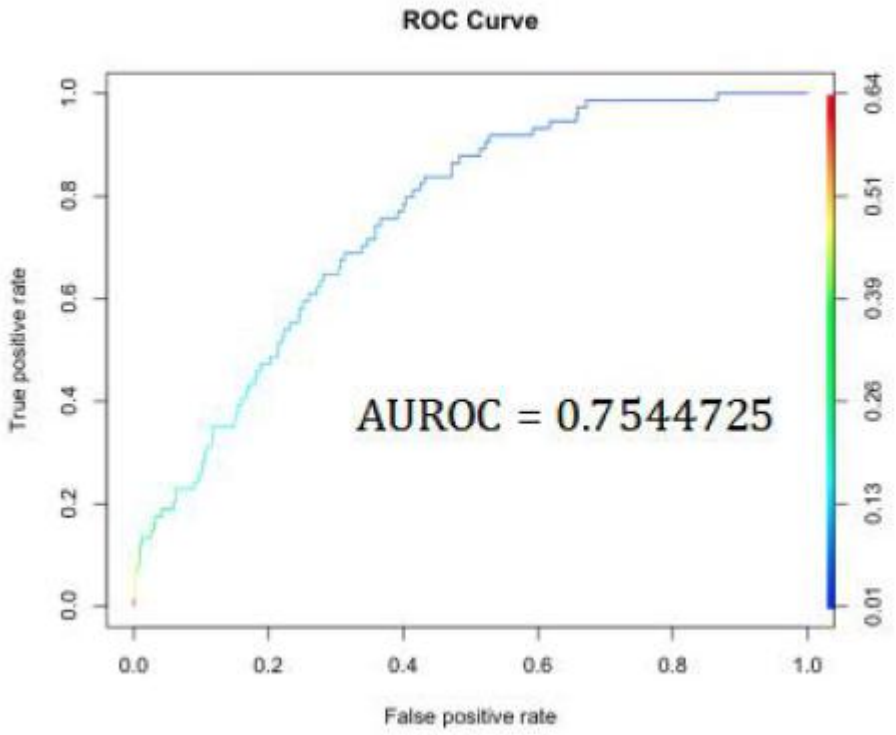
  - a measure of predictive accuracy of the logistic model

		Actual	
		0	1
Predicted	0	627	58
	1	9	13

Confusion Matrix

Since the dataset is imbalanced with proportion of events=10%, the model is able to better predict for the non occurrence of disease than its occurrence

Receiver Operating Characteristics



Mean Area for 1000 iterations: 0.746

>If the model has no predictive power, you have a 50-50 chance of correctly classifying the possibility of disease.

>More area beneath the ROC curve indicates greater predictive power Area= 0.5 for no predictive power and 1.0 for perfect predictive power

>Here, the model has a 75% chance of correct classification (quite an improvement over 50%).

04

CONCLUSION

Add your title here

- Click to add text
- Click to add text

Q&A

Thank you very much for watching

Add your title here

- Click to add text

- Click to add text