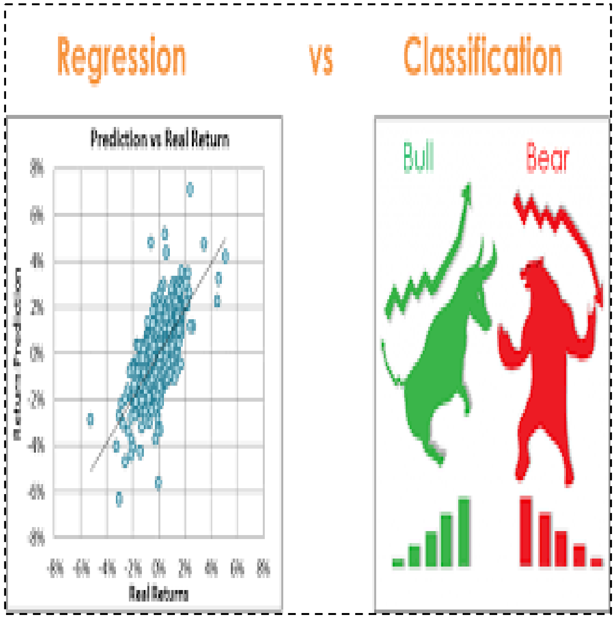
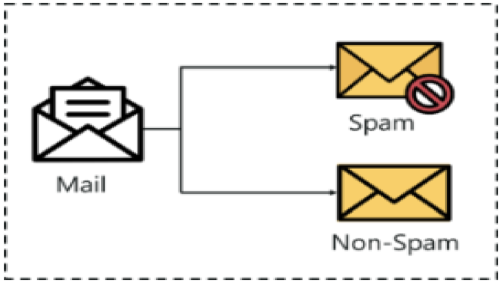
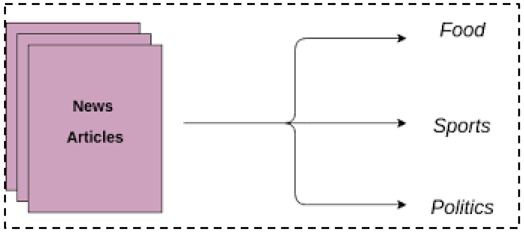
So far…

- ✓ Building Simple Linear Regression Model
- ✓ Interpretation of Regression Coefficients
- ✓ Building Multiple Linear Regression Model
- ✓ F-test, t- test and Adjusted-Rsquared
- ✓ Inclusion of Categorical Variables
- ✓ Step-Wise Regression Method in R

Classification







Classification Problems

Classification is an important category of problems in which the decision maker would like to classify the case/entity/customers into two or more groups.

Examples of Classification Problems:

- ✓ Customer profiling (customer segmentation)
- ✓ Customer Churn
- ✓ Credit Classification (low, high and medium risk)
- ✓ Employee attrition
- ✓ Fraud (classification of transaction to fraud/no-fraud)
- ✓ Stress levels
- ✓ Text Classification (Sentiment Analysis)
- ✓ Outcome of any binomial and multinomial experiment

Classification Algorithms

Logistic Regression

Discriminant Analysis

Decision Tree

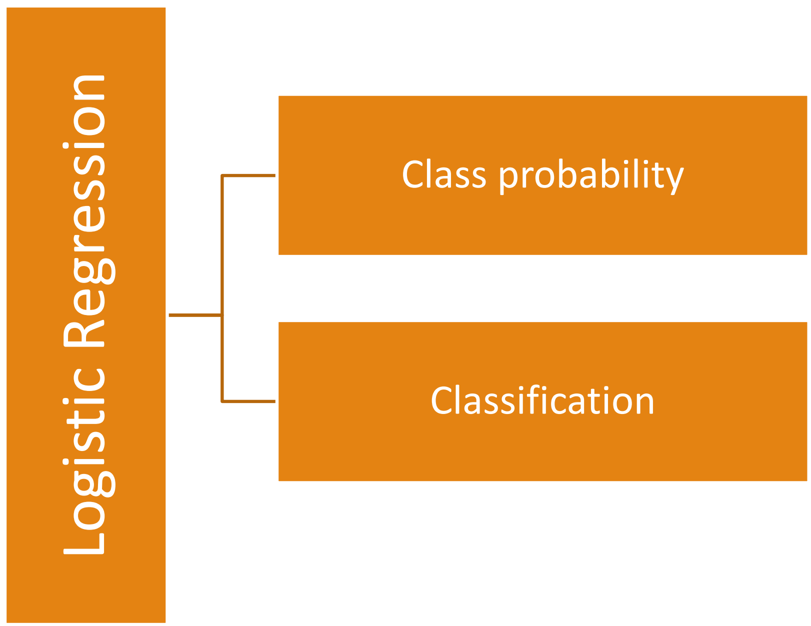
Ensemble Method

Naïve Bayes

Support Vector Machines method

Other methods

Logistic Regression


Logistic Regression
Class probability
Classification

Logistic v/s Linear

Linear regression not appropriate where response is categorical

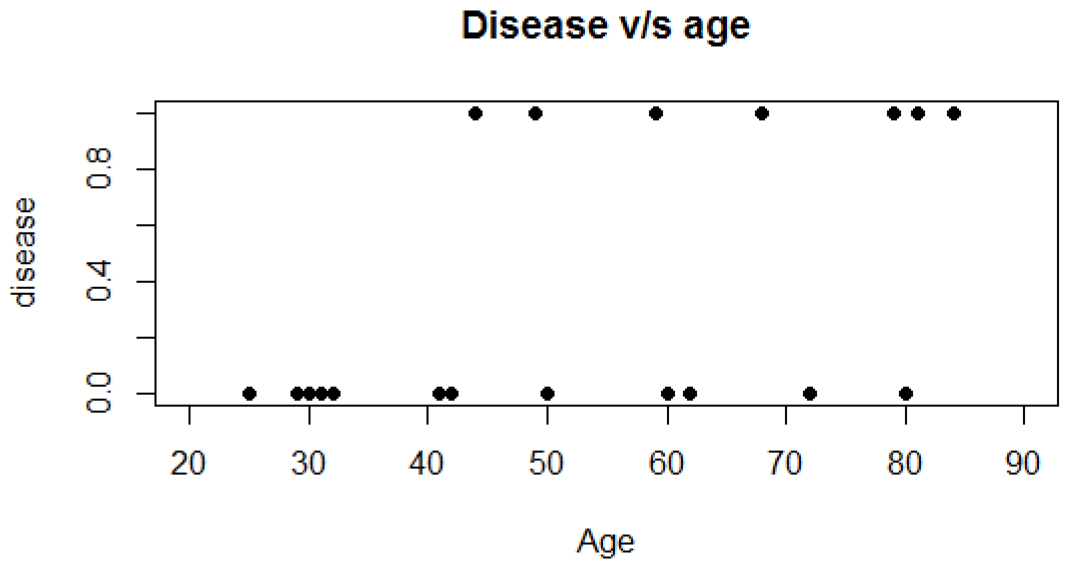
Alternatively, Logistic Regression method describes relationship between categorical response and set of predictors

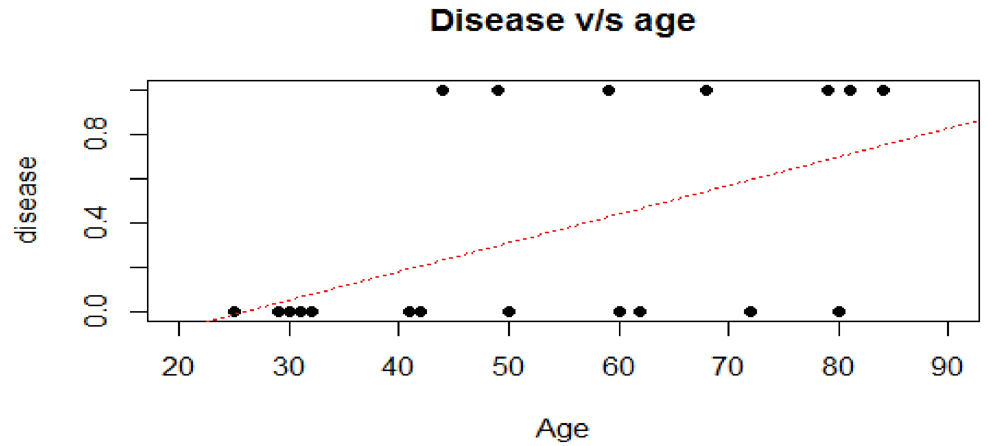
Specifically, we explore applications with dichotomous response

Example: Suppose researchers interested in potential relationship between patient *age* and presence/absence of *disease*

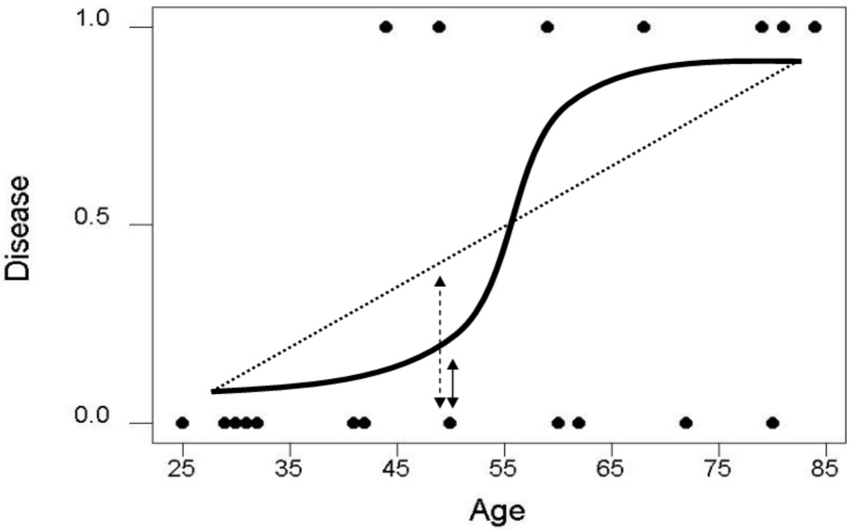
Data set includes 20 patients

	age	disease
1	25	0
2	29	0
3	30	0
4	31	0
5	32	0
6	41	0
7	41	0
8	42	0
9	44	1
10	49	1
11	50	0
12	59	1
13	60	0
14	62	0
15	68	1
16	72	0
17	79	1
18	80	0
19	81	1
20	84	1

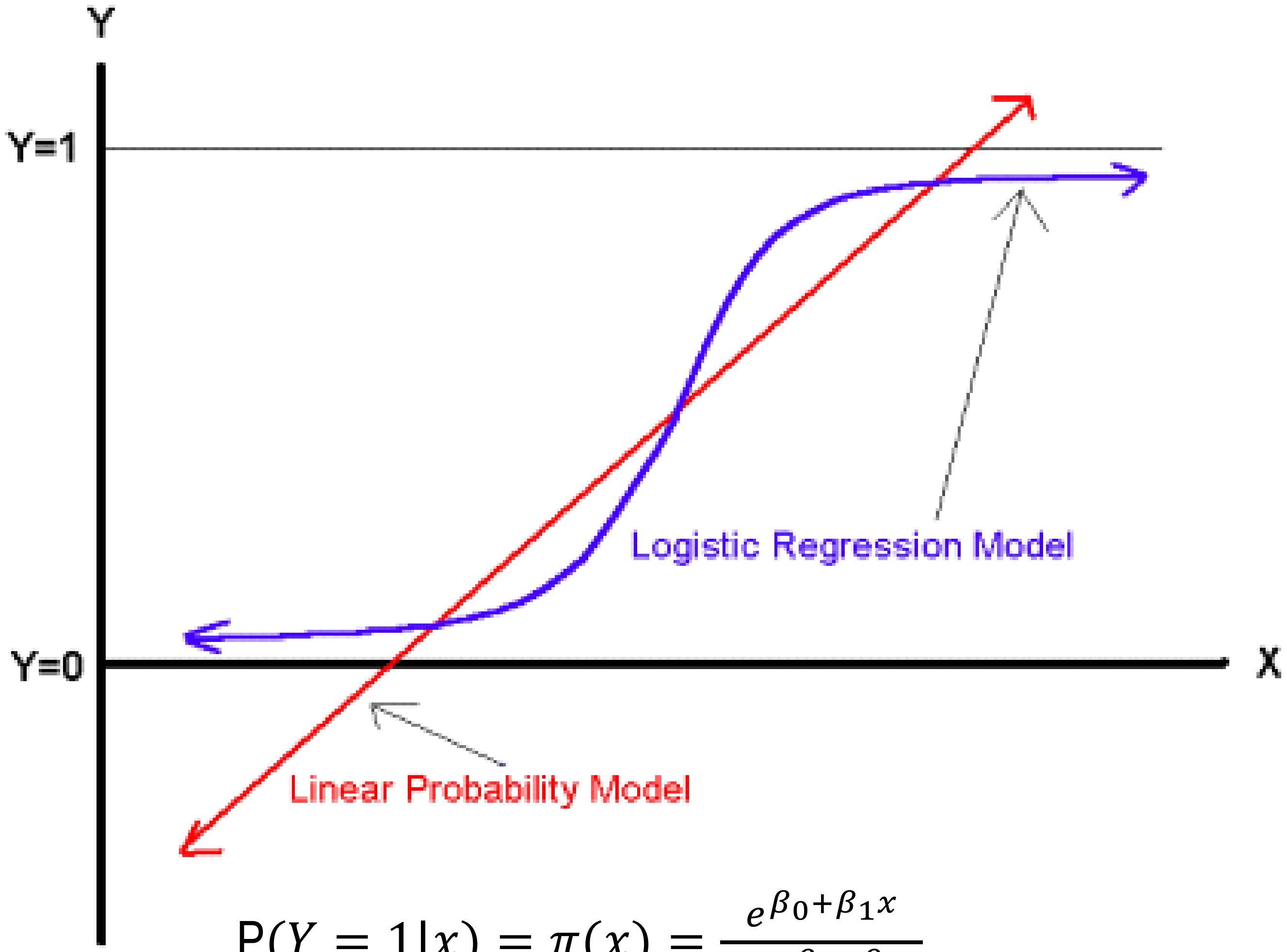




Plot shows least squares regression line (straight) and logistic regression line (curved) for *disease* on *age*

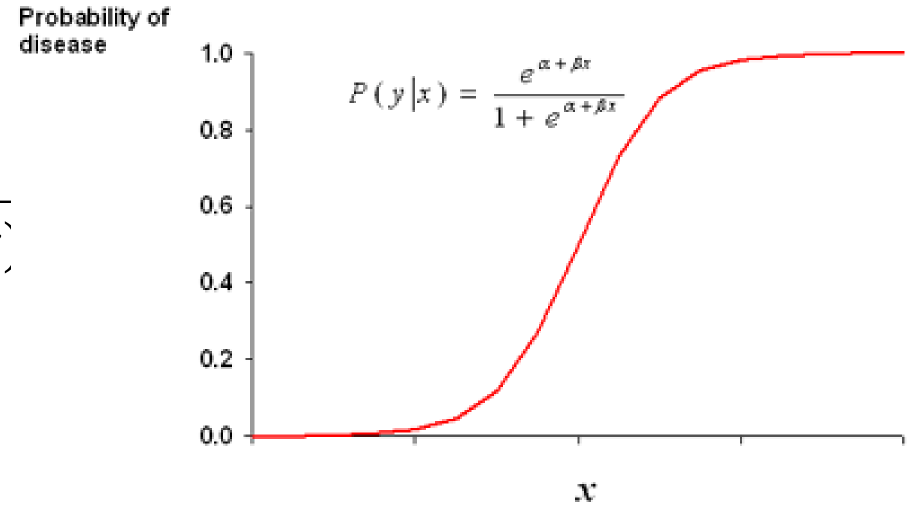


Comparing the LP and Logit Models



$$P(Y = 1|x) = \pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

$$P(Y=1 \mid X=x) = \pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$



Estimating β with b

- b = 0 implies that P(Y|x) is same for each value of x
- b > 0 implies that P(Y|x) is increases as the value of x increases
- b < 0 implies that P(Y|x) is decreases as the value of x increases

Interpreting Logistic Regression Output

$$\hat{g}(x) = -4.372 + 0.06696(50) = -1.024$$

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} = \frac{e^{-4.372 + 0.06696\ (age)}}{1 + e^{-4.372 + 0.06696\ (age)}}$$

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} = \frac{e^{-1.024}}{1 + e^{-1.024}} = 0.26$$

```
Call:
glm(formula = disease ~ age, family = binomial, data = patients)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6136  -0.6591  -0.4310   0.7856   1.8118

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.37210    1.96555  -2.224   0.0261 *
age          0.06696    0.03223   2.077   0.0378 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.898  on 19  degrees of freedom
Residual deviance: 20.201  on 18  degrees of freedom
AIC: 24.201

Number of Fisher Scoring iterations: 4
```

Estimate probability of *disease* present in particular patient, *age* = 50

```
Call:
glm(formula = disease ~ age, family = binomial, data = patients)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6136  -0.6591  -0.4310   0.7856   1.8118

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.37210    1.96555  -2.224   0.0261 *
age          0.06696    0.03223   2.077   0.0378 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.898  on 19  degrees of freedom
Residual deviance: 20.201  on 18  degrees of freedom

Number of Fisher Scoring iterations: 4
```

Deviance is a measure of badness of fit. Higher numbers indicate worse fit.

The null deviance is the deviance when the response variable is predicted by a model that includes only the intercept (grand mean).

The Residual Deviance is the deviance when the response variable is predicted by the predictor variable(s).

Overall Model Significance

$H_0 : \beta_1 = \beta_2 \ldots = \beta_k = 0;\ H_A :$ Not all $\beta_i = 0$

Test statistic:

Null Deviance (Model with no predictor)- Residual Deviance(Model with all predictors)

It follows a chi-square distribution with k degree of freedom

P(chi-square > Test statistic)

0.01699968

Inference: Is a predictor Significant?

Wald test: hypothesis test for assessing significance of predictor

Under Ho: that $\beta_1 = 0$, Ha: $\beta_1 \neq 0$

Z_{Wald} statistic follows standard normal distribution:

$$Z_{Wald} = \frac{b_1}{SE(b_1)}$$

```
Call:
glm(formula = disease ~ age, family = binomial, data = patients)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6136  -0.6591  -0.4310   0.7856   1.8118

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.37210    1.96555  -2.224   0.0261 *
age          0.06696    0.03223   2.077   0.0378 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.898  on 19  degrees of freedom
Residual deviance: 20.201  on 18  degrees of freedom
AIC: 24.201

Number of Fisher Scoring iterations: 4
```

	npreg	glu	bp	skin	bmi	ped	age	type
1	6	148	72	35	33.6	0.627	50	Yes
2	1	85	66	29	26.6	0.351	31	No
3	1	89	66	23	28.1	0.167	21	No
4	3	78	50	32	31	0.248	26	Yes
5	2	197	70	45	30.5	0.158	53	Yes
6	5	166	72	19	25.8	0.587	51	Yes
7	0	118	84	47	45.8	0.551	31	Yes

npreg: number of pregnancies.
glu: plasma glucose concentration in an oral glucose tolerance test.
bp: diastolic blood pressure (mm Hg).
skin: triceps skin fold thickness (mm).
bmi: body mass index (weight in kg/(height in m)\^2).
ped: diabetes pedigree function.
age: age in years.
Type: Yes or No, for diabetic according to WHO criteria

Refer to file Pima.te.

Q.1 Check the overall significance of the model. What is the value of statistic and what is its degree of freedom

Q.2 Which all variables are significant? What is the z-statistic for bp, skin?

Q.3 Predict the probability of disease of the person with the values

(npreg=2,glu= 90,bp=70,bmi=44,ped=0.487,age=60)

Q.4 Whether the person is diabetic or not.

Cutoff Probability

$$Y = 1; if\ P(y = 1) \geq a$$

$$Y = 0; if\ P(y = 1) < a$$

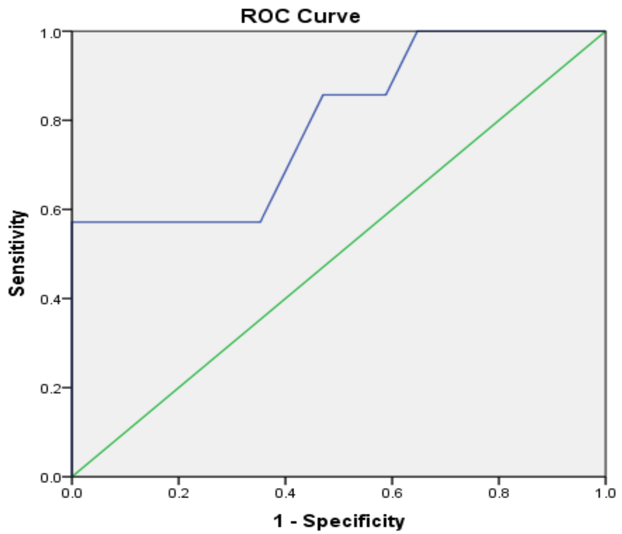
Model Evaluation

CONFUSION MATRIX	Predicted Positive	Predicted Negative
Actual Positive	True Positives (TP)	False Negatives (FN)
Actual Negative	False Positives (FP)	True Negatives (TN)

Sensitivity= $\frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$

Specificity= $\frac{\text{True Negative (TN)}}{\text{True Negative (TN)} + \text{False Positive (FP)}}$

Accuracy= $\frac{\text{True Positive (TP)} + \text{True Negative } (TN)}{Total}$



Area under ROC curve

	Predicted Value	
ActualValue	0 (negative)	1 (positive)
0	40	4
1	14	9

False positive		4
False Negative		14
True Positive		9
True Negative		40
Accuracy		49/67
Sensitivity		9/23
Specificity		40/44
TPR		9/23
FPR		4/44

Accuracy Paradox

	Predicted Value	
ActualValue	0 (negative)	1(positive)
0	80(TN)	0 (FP)
1	20 (FN)	0 (TP)

- accuracy <- (p[1,1]+p[2,2])/sum(p)
- 0.8

- Suppose we have data set with 100 observations with 20 observations with value "1" and 80 with "0".
- A classifier predicted all the observations to be "0"
- Accuracy will be 0.8

Youden's Index

Sensitivity+ Specificity -1

Optimal Cutoff