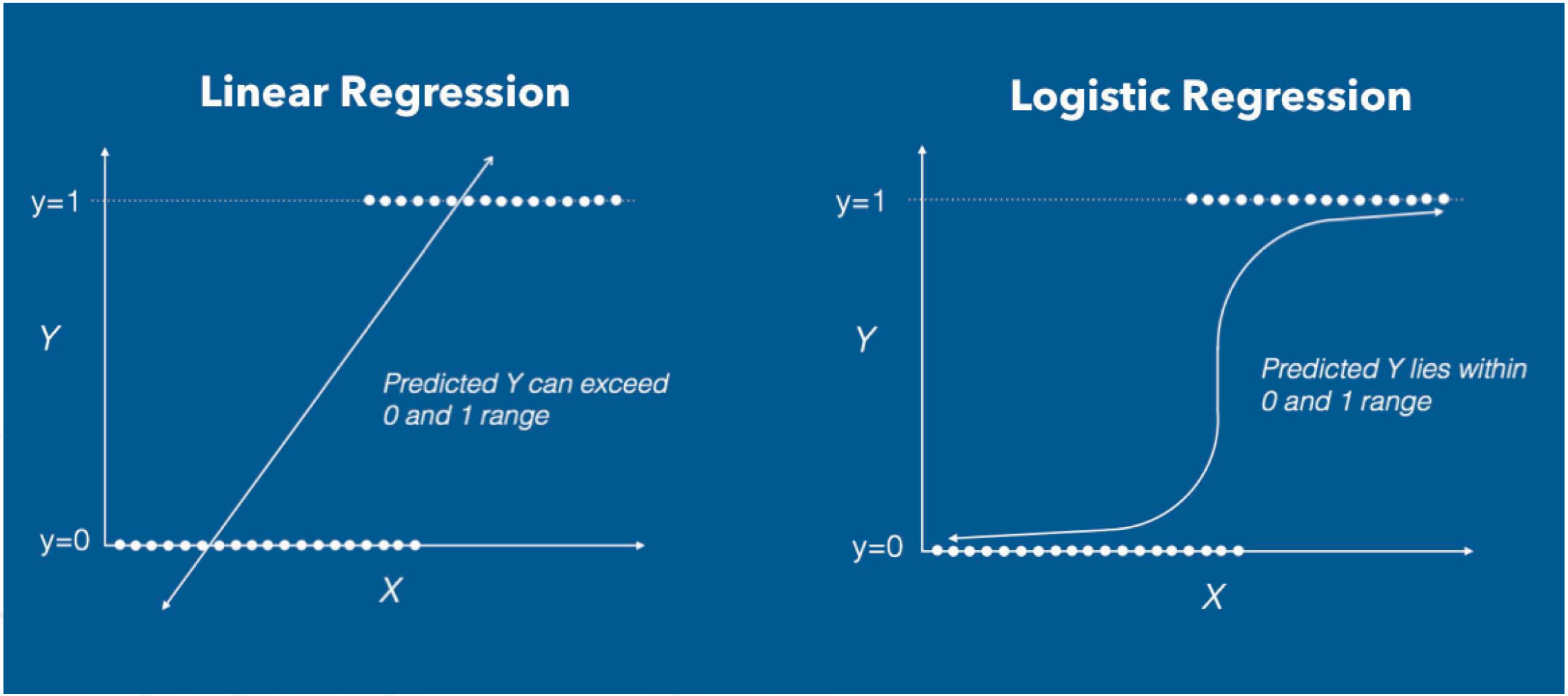
Logistic Regression Model

王　鹏

201911



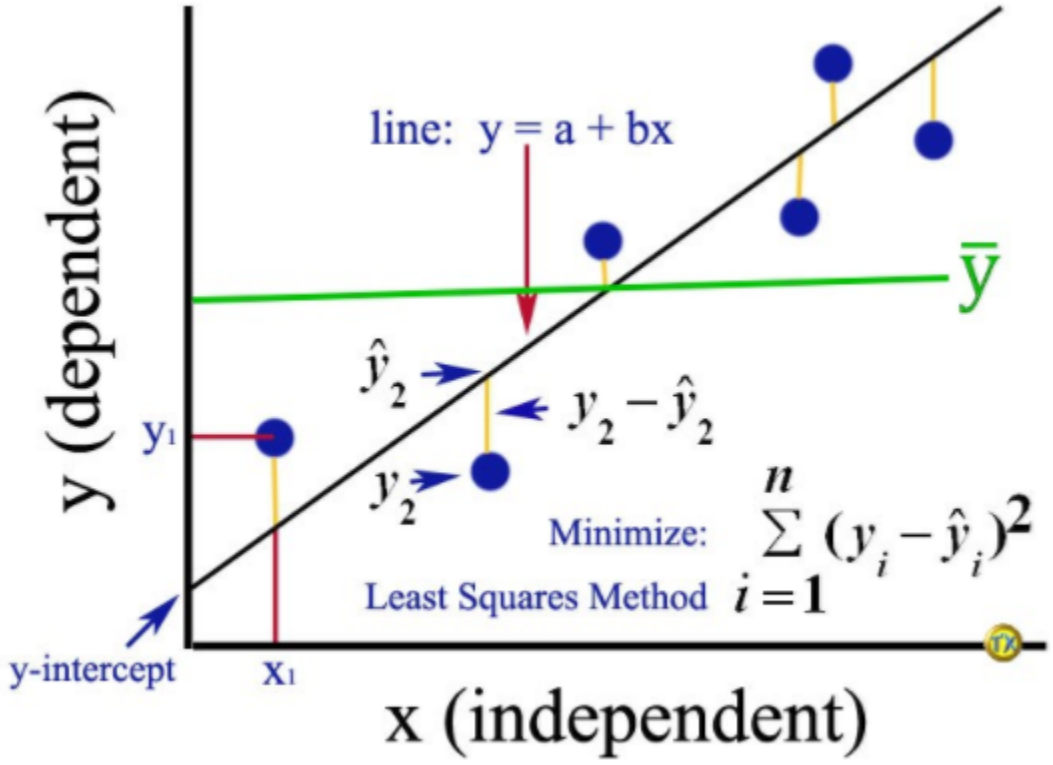
$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

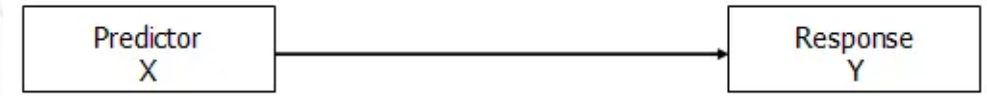
$$Y = \ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

https://www.machinelearningplus.com/machine-learning/logistic-regression-tutorial-examples-r/

Regression analysis

回归分析（Regression Analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法





Step1：计算中心点（x和y的平均值）

Step2：线性拟合（最小二乘）

Step3：计算斜率b、截距a和决定系数R^2

$y_i \qquad \bar{y} \qquad \hat{y} \qquad a = \bar{y} - b\bar{x}$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\text{F-statistics} = \frac{SSR/(p-1)}{SSE/(n-p)}$$

$SSE = \Sigma(y_i - \hat{y}_i)^2$	SSE = sum of squares due to error
$SST = \Sigma(y_i - \bar{y})^2$	SST = total sum of squares
$SSR = \Sigma(\hat{y}_i - \bar{y})^2$	SSR = sum of squares due to regression

$$R^2\_adjusted = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

$$R^2 = \frac{SSR}{SST}$$

Novogene 诺禾致源

Providing advanced genomic solutions!

What is odds and odds ratio (OR)

Odds are defined as the ratio of the probability of Y (p) and the probability of N (1-p)

$$Odds = \frac{p}{1-p}$$

Y: p = 0.8; N: 1-p=0.2;

Odds(Y) = 0.8/0.2 = 4

Odds(N) = 0.2/0.8 = 0.25

Odds ratio (OR)

OR = 4 / 0.25 = 8

Suppose that p is the probability of smoking (S), 1-p is of no smoking (N), here is the table

	D	H
Smoking (S)	8	3
No smoking (N)	2	7

S: p = 0.8; N: 1-p=0.2;

Odds(S) = 0.8/0.2 = 4

Odds(N) = 0.2/0.8 = 0.25

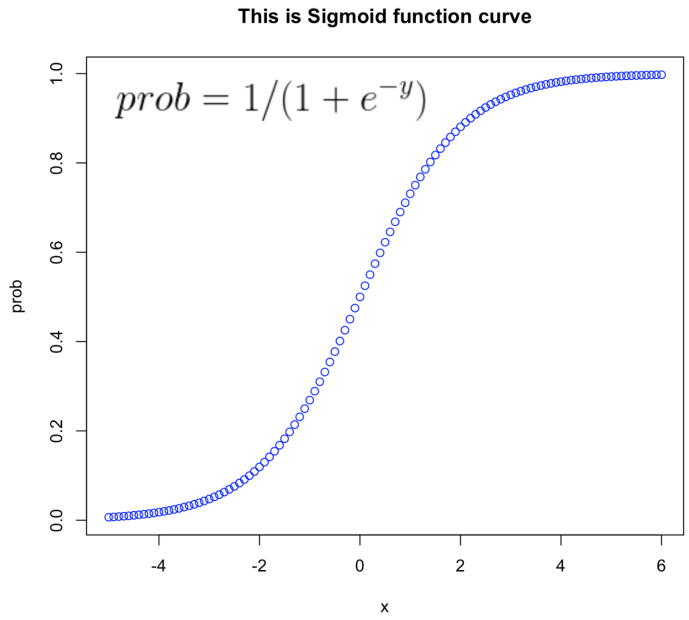
For smoking, the odds ratio of being disease is 8.

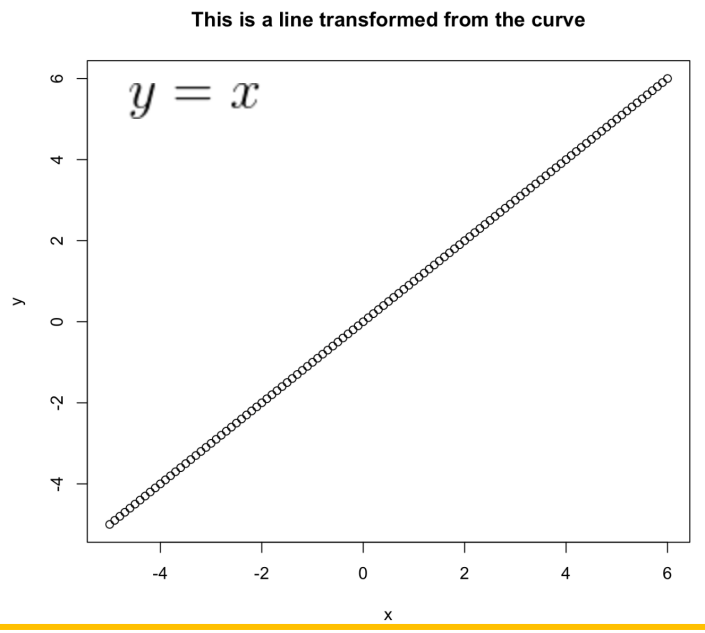
The factor of smoking have a big effect on health status.

https://stats.idre.ucla.edu/stata/faq/how-do-i-interpret-odds-ratios-in-logistic-regression/

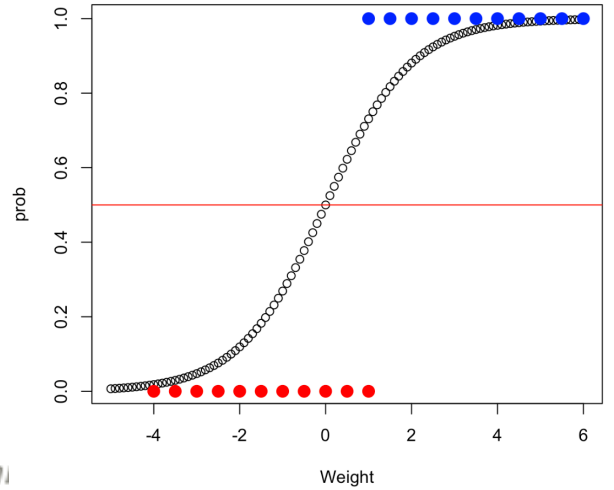
The relationship between S curve and straight line

```
###x limits
x <- seq(-5, 6, 0.1)
###sigmoid function
###y = log(p/(1-p)) = x
###p = exp(x) / (1 + exp(x))
prob <- exp(x) / (1 + exp(x))
plot(x, prob, col='blue')
###curve changes to straight line
y <- log(prob/(1-prob))
plot(x, y)
```

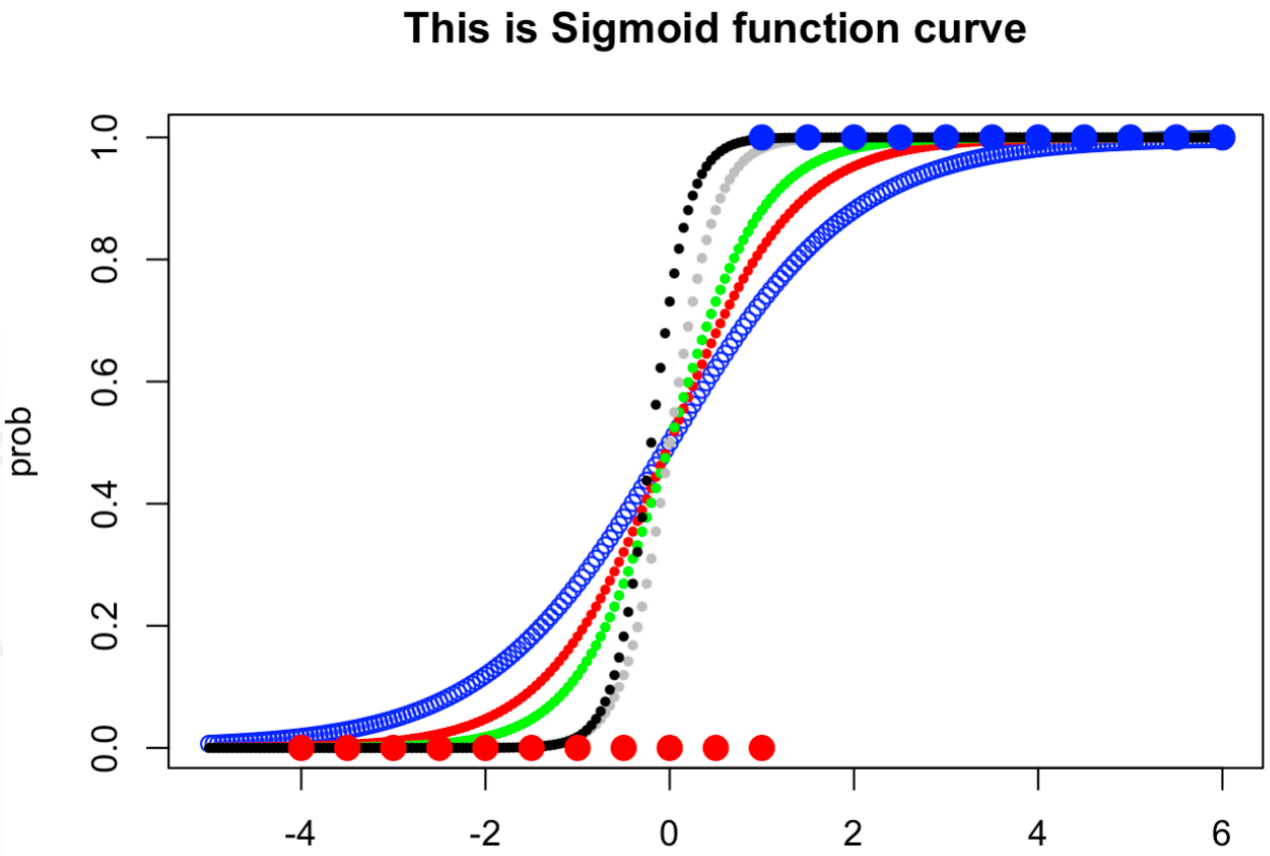





```
x0 <- seq(-4, 1, 0.5)
x1 <- seq(1, 6, 0.5)
plot(x, prob, col='black', xlab = 'Weight')
points(x0, rep(0, length(x0)), pch = 19, cex = 1.5, col = 'red')
points(x1, rep(1, length(x0)), pch = 19, cex = 1.5, col = 'blue')
```



How to determine the best fitted curve



There are multiple curves to fit the points,

Which is the best fitted "S" curve?

Maximum Likelihood (ML) or Log(ML)

$$ML = P_1.fit \cdot P_2.fit \cdot \ldots P_n.fit = \prod P_i.fit$$

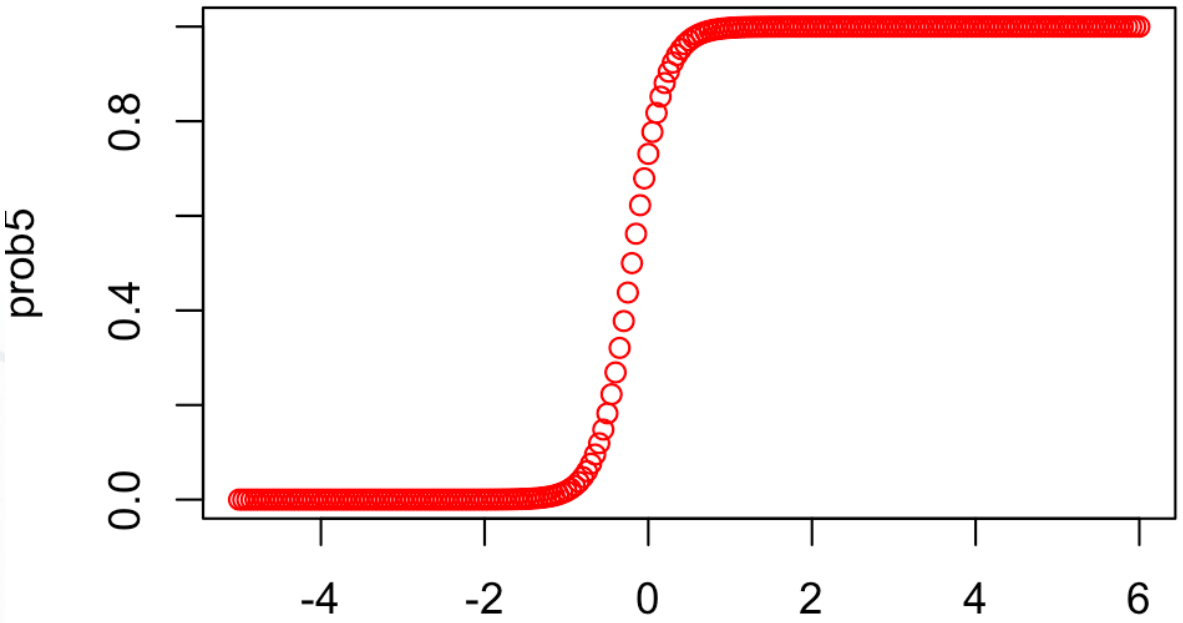
$$log(ML) = \prod log(P_i.fit)$$

ML of one curve reach the maximum, that's it.

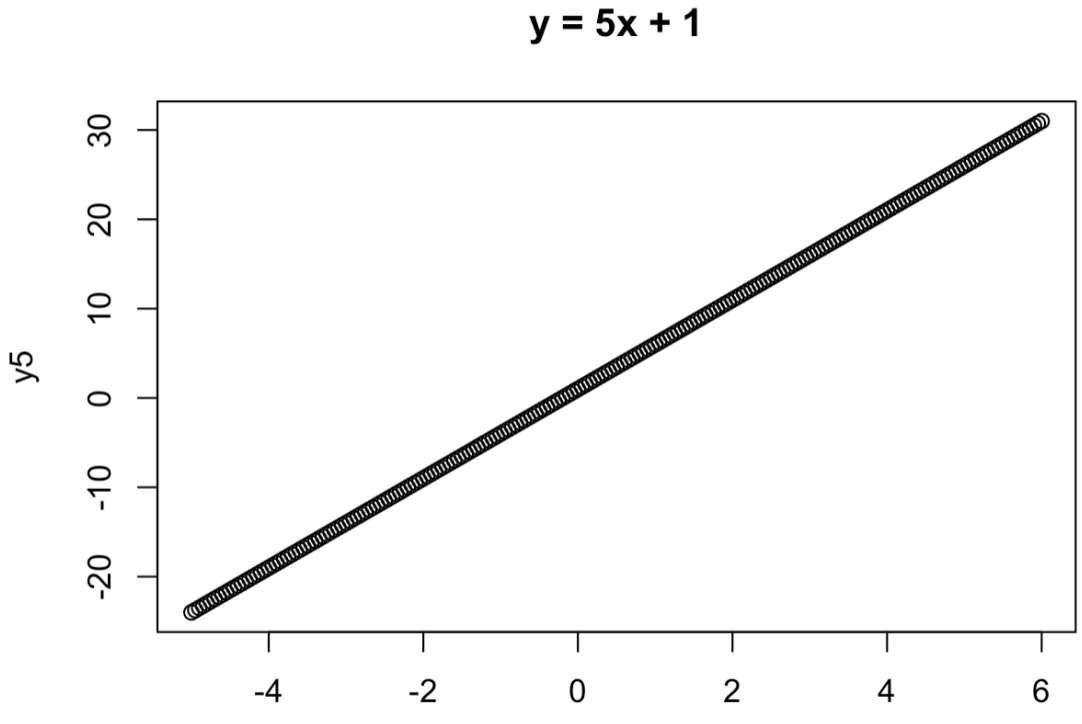
The parameters of LR

Log odds

$$p = \exp(5x+1)/(1+\exp(5x+1))$$
$$y=\log(p/(1-p)) = 5x+1$$



Logistic regression curve



transformed regression line

Coefficient and intercept

Explanation of LR using R step by step

1 Data type

	group	Lactobacillus	Helicobacter	Epulopiscium	Hannaella	Coxiella	Esch
L1DC09M	L1DCM	1.99335940	0.09815416	-1.54406781	-1.71797865	-0.72953619	-0.3
L1DC10M	L1DCM	-0.64490809	-1.07356616	1.72948022	1.19722809	-0.13185950	-0.5
L1DC11M	L1DCM	0.67170367	-1.00006004	0.28043756	0.19714113	-0.81775153	-0.5
L1DC12M	L1DCM	2.31214745	0.02017730	1.87549458	-1.78145730	-0.07415404	-0.6
L1DC13M	L1DCM	-0.64687772	-0.81067915	1.23096551	0.79821409	1.32689085	-0.3
L1DC14M	L1DCM	-0.61021647	1.90734016	-0.08928165	-0.33995523	0.34485666	-0.0
H2DC01M	H2DCM	-0.61731748	0.41818697	-1.27229935	-1.53072848	-0.92565818	-0.5
H2DC02M	H2DCM	-0.24745734	1.01031926	-0.27780547	-0.14760138	0.66009316	-0.4
H2DC06M	H2DCM	-0.45946103	2.31458317	-0.97496714	-1.05107625	-0.80681919	-0.5
H2DC07M	H2DCM	2.12216802	-0.99222776	-0.90363036	-0.66105233	-1.01615422	-0.6

Known group level

Features (various taxa)

Groups must be factors

Explanation of LR step by step

2 Model fit

```
lrm <- glm(group ~ Lactobacillus, data=dt, family="binomial"); summary(lrm)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.002613   0.451597   0.006    0.995
Lactobacillus 0.285924   0.475429   0.601    0.548
```

Feature selection

Intercept: when Lactobacillus is 0, the log odds of L1DCM is 0.0026

Coefficients: when Lactobacillus change one unit, the log odds of L1DCM increase 0.286

zvalue and **Pr** is the Wald test statistics and p value

Odds ratio can be as e^0.286 = 1.33

One unit of Lactobacillus changes, the odds of L1DCM increase 1.33

Explanation of LR step by step

3 Model test and 4 using model

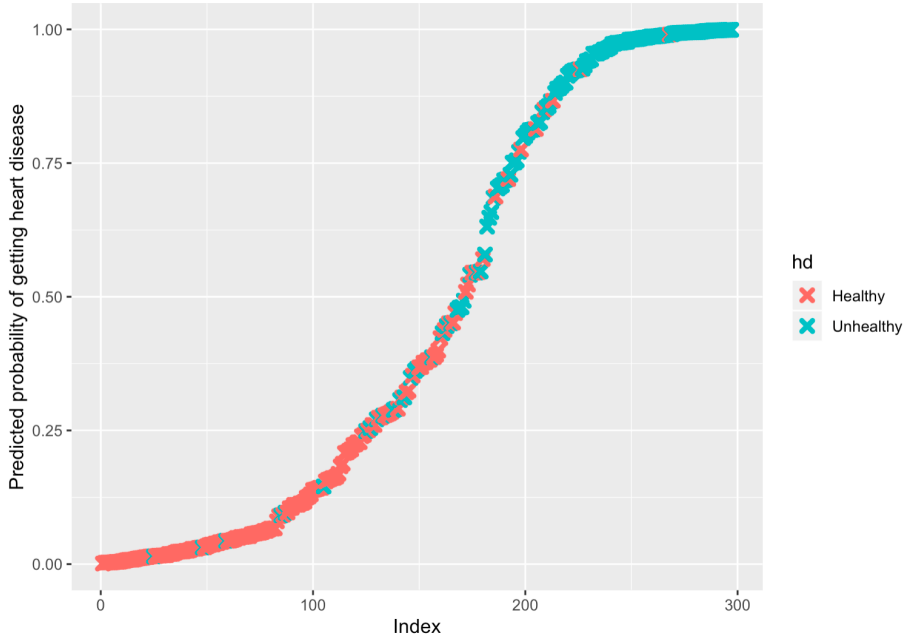
```
sqr <- (lrm$null.deviance/-2 - lrm$deviance/-2) / (lrm$null.deviance/-2)
chisq.value <- 2*(lrm$deviance/-2 - lrm$null.deviance/-2)
chisq.pval <- with(lrm, pchisq(null.deviance - deviance, df.null -
df.residual, lower.tail = F))
```

sqr is the R^2, which is the goodness of this model.

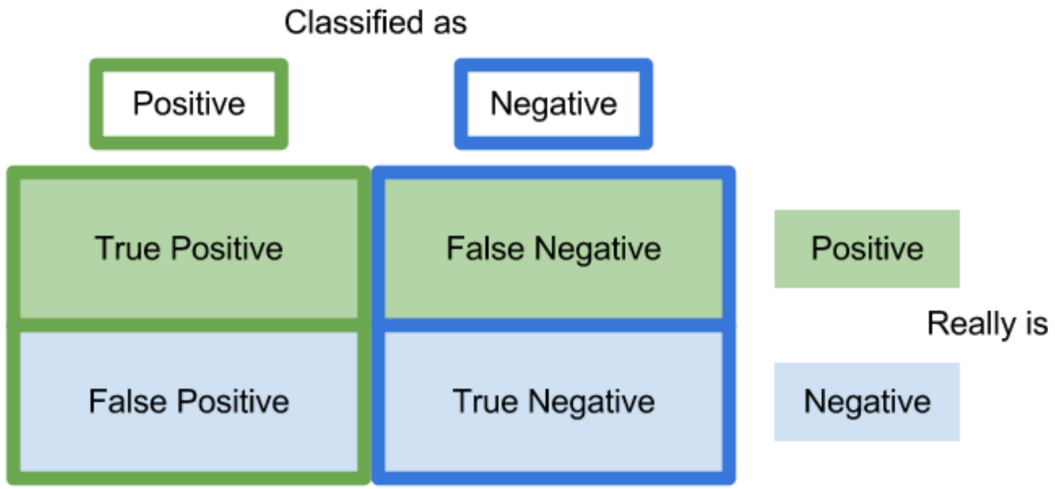
To test the robustness of the model, Chisq test was performed, and getting the **chisq.value** and **chisq.pval**

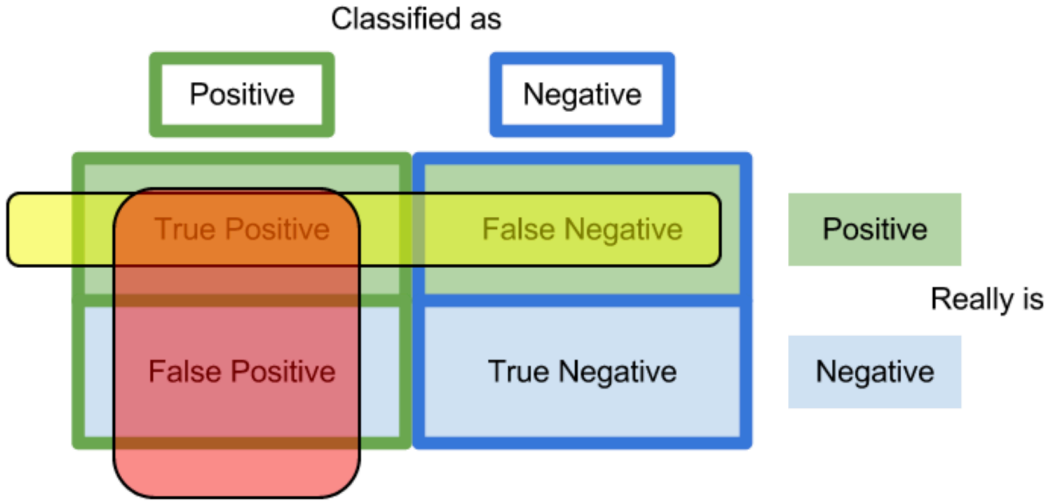
https://statquest.org/2018/07/23/statquest-logistic-regression-in-r/

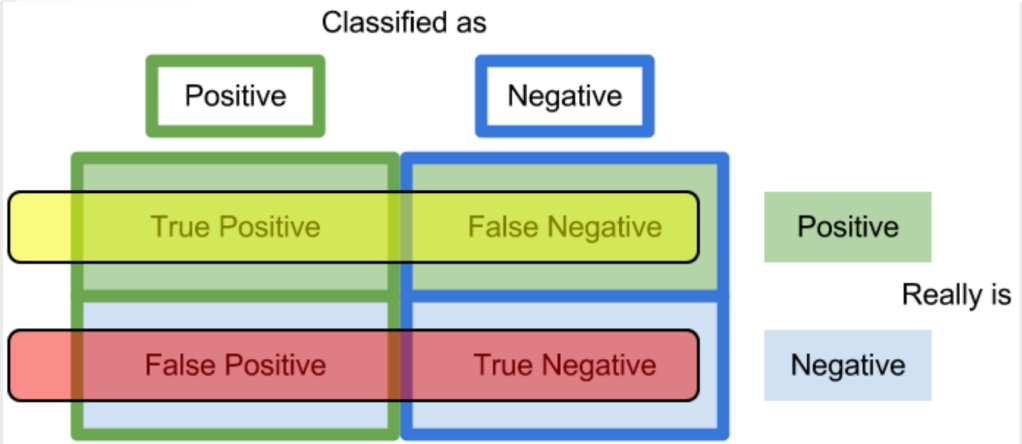
Using function `predict.glm()`, we can (i) get the model accuracy, and (ii) predict the class of new samples



Precision, recall, sensitivity and specificity







Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

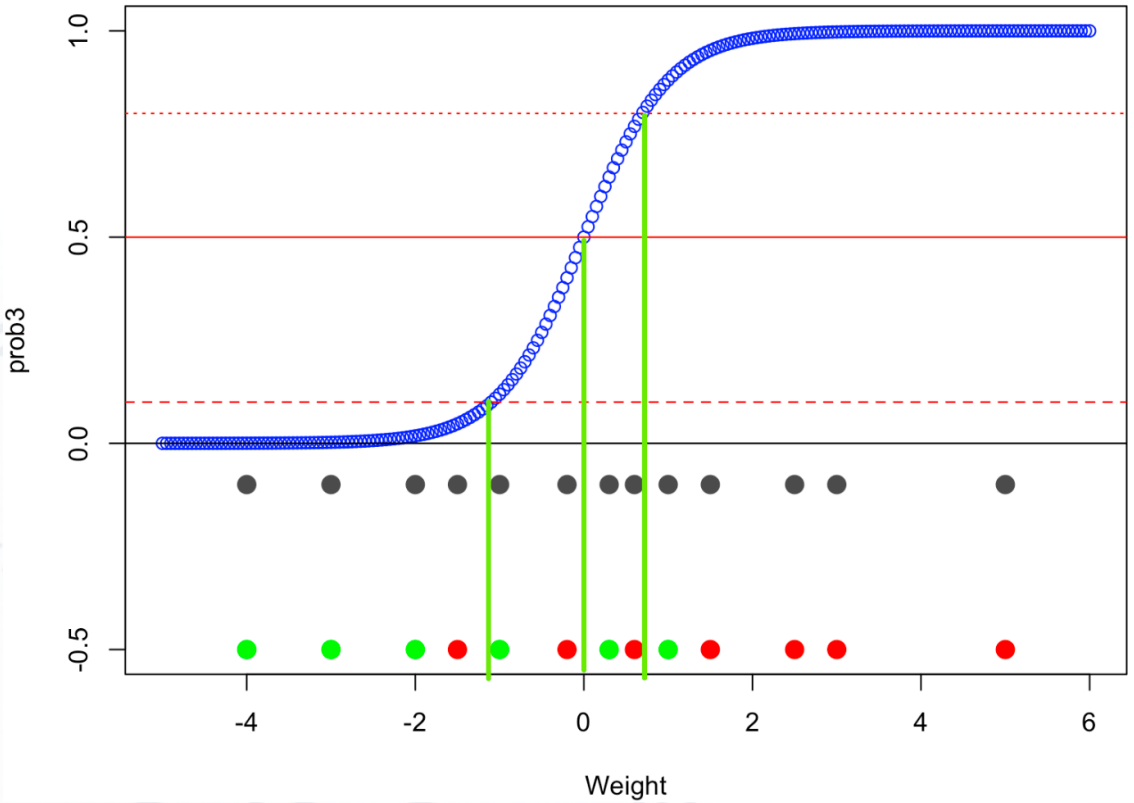
Recall = TP / (TP + FN)

Precision = TP / (TP + FP)

https://uberpython.wordpress.com/2012/01/01/precision-recall-sensitivity-and-specificity/

ROC and AUC

Receiver Operating Characteristics (ROC) and Area Under The Curve (AUC)



https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

cutoff=0.1		Predict	
		P(red)	N(green)
Actual	P(red)	6	1
	N(green)	3	3

Sensitivity = TP / (TP + FN) = 6 / (6 + 1) = 0.857

Specificity = TN / (TN + FP) = 3 / (3 + 3) = 0.500

cutoff=0.5		Predict	
		P(red)	N(green)
Actual	P(red)	5	2
	N(green)	2	4

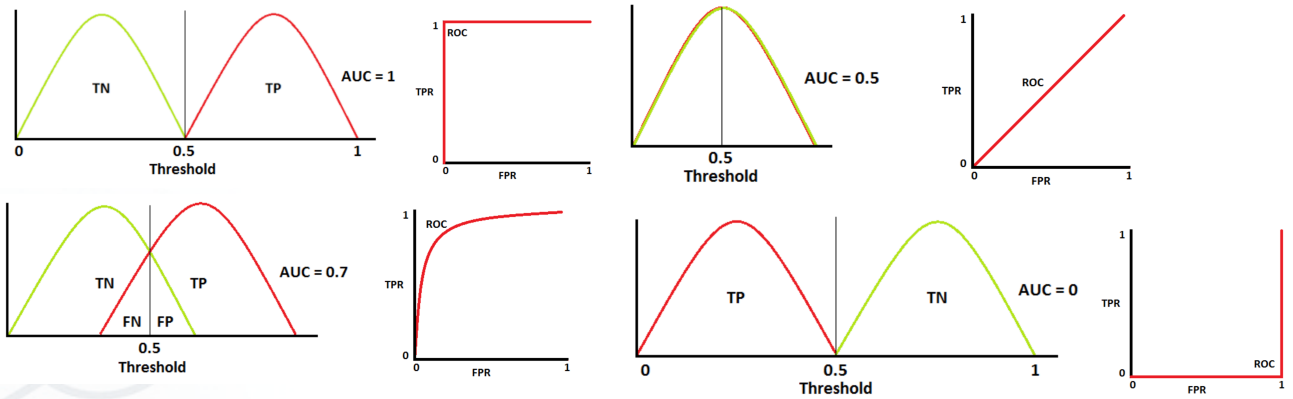
Sensitivity = TP / (TP + FN) = 5 / (5 + 2) = 0.857

Specificity = TN / (TN + FP) = 4 / (4 + 2) = 0.667

cutoff=0.8		Predict	
		P(red)	N(green)
Actual	P(red)	4	3
	N(green)	1	5

Sensitivity = TP / (TP + FN) = 4 / (4 + 3) = 0.571

Specificity = TN / (TN + FP) = 5 / (1 + 5) = 0.833



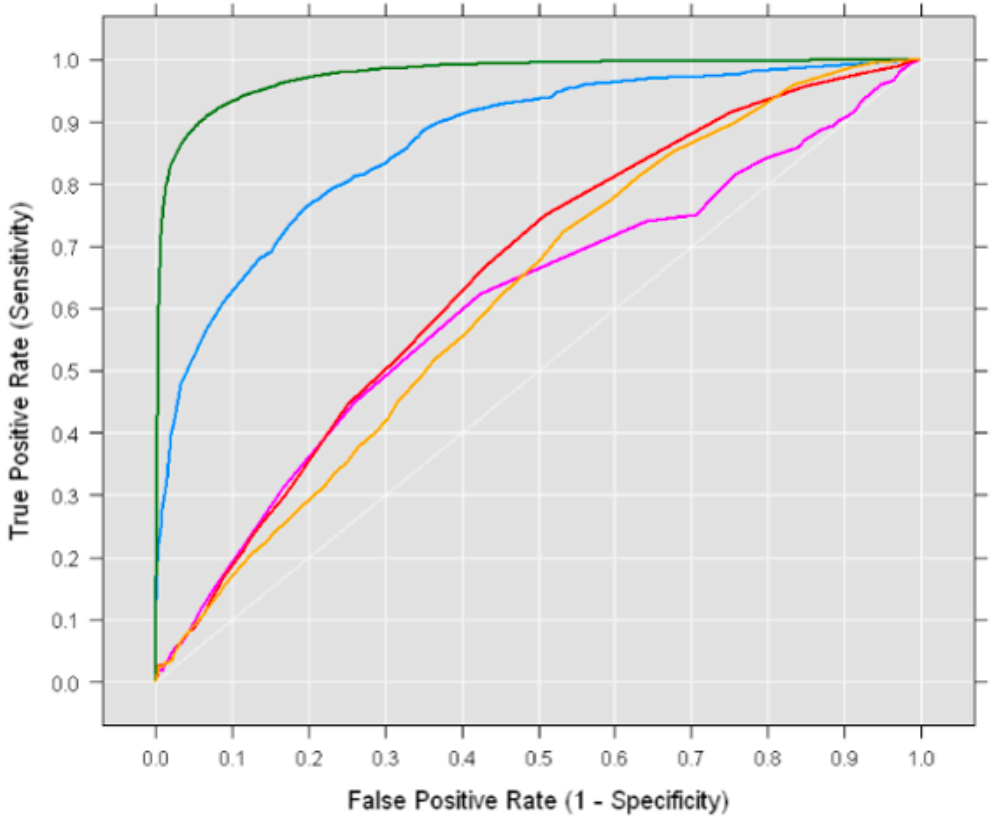
Novogene 诺禾致源

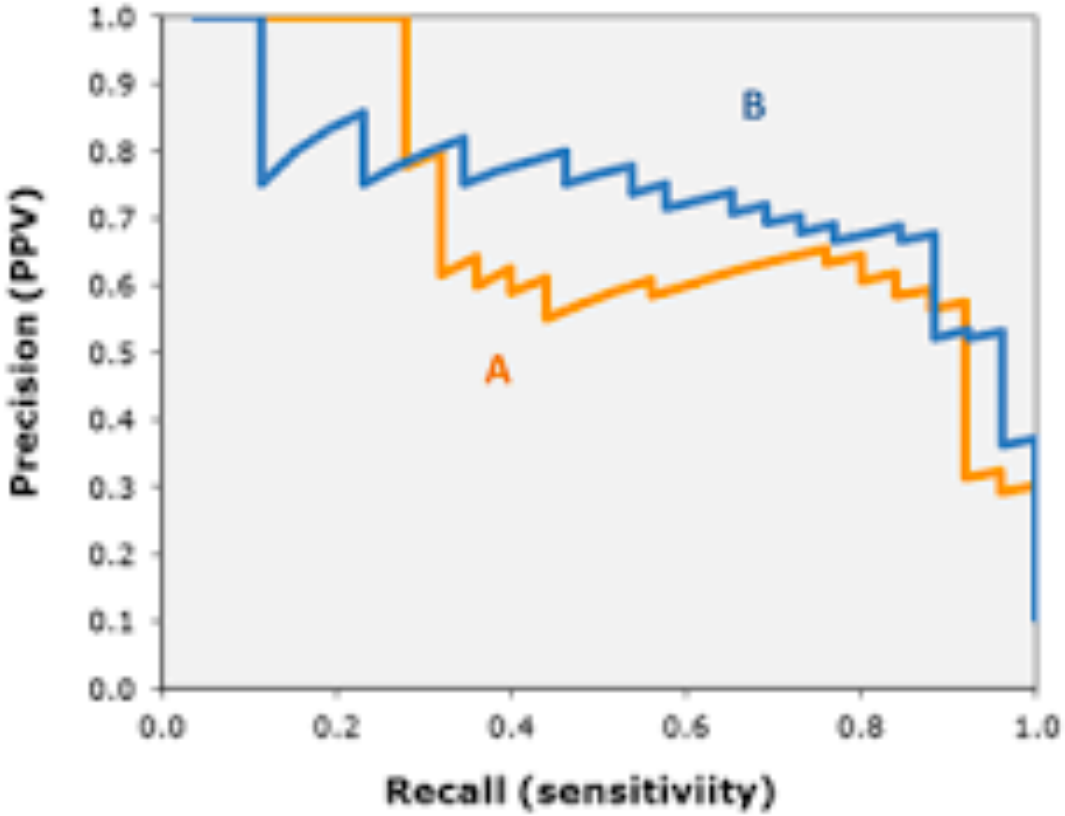
Providing advanced genomic solutions!

ROC AUC and Precision-Recall curve

ROC can be plotted according to the Sensitivity and Specificity or 1- Specificity

AUC is the area under the curve





https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/

Novogene 诺禾致源

Providing advanced genomic solutions!

Thanks for your attention!

更多关注，敬请留意：www.novogene.cn