Chapter 9: Multiple and logistic regression

Wen-Han Hwang
(Slides primarily developed by Mine Çetinkaya-Rundel from OpenIntro.)

Institute of Statistics
National Tsing Hua University
Taiwan

OpenIntro

Outline

Introduction

Regression so far ...

At this point we have covered:

$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

$\Rightarrow Y|X \sim$ Normal Dist

- Simple linear regression
  - Relationship between numerical response and a numerical or categorical predictor
- Multiple regression
  - Relationship between numerical response and multiple numerical and/or categorical predictors

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

However, several challenges remain:

- **Complex Predictors:** Handling predictors with nonlinear relationships or intricate dependency structures.
- **Diverse Response Types:** Addressing different types of response variables, such as binary, percentage, categorical, and count data.

$0 \le Y \le 1$ $\quad Y = \{A, B, C\}$ $\quad Y = 0, 1, 2, 3, \ldots$ $\quad Y = \{0, 1$

Logistic Regression: Specifically focuses on cases with binary response variables.

Understanding Odds

Odds provide another method for quantifying the likelihood of an event and are frequently used in contexts like gambling and logistic regression.

Definition:

- For an event E, the odds are defined as:

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

- If the odds of E are given as x to y, then:

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

which leads to:

$$P(E) = \frac{x}{x+y}, \quad P(E^c) = \frac{y}{x+y}$$

Example - Donner Party

In 1846, the Donner and Reed families departed from Springfield, Illinois, bound for California in covered wagons. By July, the group, known as the Donner Party, had reached Fort Bridger, Wyoming. Here, they opted to try a new, untested route to the Sacramento Valley. The party had grown to 87 people and 20 wagons.

Their journey was hampered by difficult crossings of the Wasatch Range and the desert west of the Great Salt Lake. They became stranded in the eastern Sierra Nevada mountains due to heavy snows in late October. When rescuers arrived on April 21, 1847, only 47 of the original 87 members had survived, the rest succumbing to famine and extreme cold.

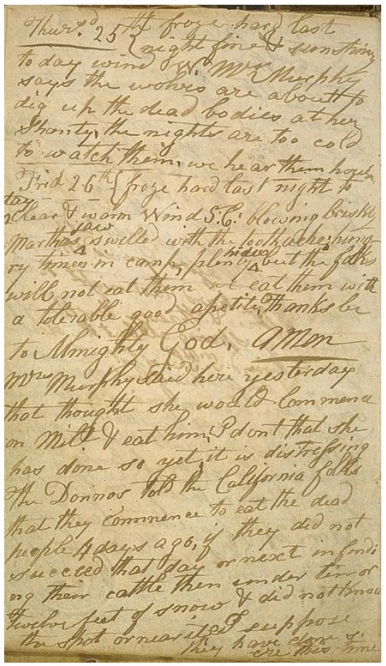
Source: Ramsey, F.L. and Schafer, D.W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis* (2nd ed.)

唐納大隊（Donner Party），又稱作**唐納-瑞德大隊**（Donner-Reed Party），指的是一群在1846年春季由美國東部出發，預計前往加州的移民隊伍，他們是由數個家庭組成的篷車大隊。由於錯誤的資訊，他們的旅程遭受延遲，導致他們在1846年末到1847年初之間受困在內華達山區度過寒冬。在惡劣的環境下，接近半數成員遭到凍死或者餓死，部分生存者依靠食人存活下來。絕大多數受害者是因疾病、飢餓等原因自然死亡後被食用，但有兩位原住民嚮導是被故意殺害以而成為被食用的受害者。

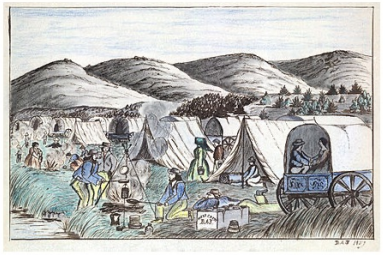
前往西部的旅途通常需要費時四到六個月，但唐納大隊採取了一條被稱為黑斯廷近道的新路徑，他們橫越猶他州的瓦薩奇山脈及大鹽湖沙漠，接著來到洪堡河谷，到達現今的內華達州，此時他們已經失去了眾多的牲畜，隊伍內部也產生分裂。

在1846年11月初，他們終於來到最後的關卡，內華達山脈，然而一場提早到來的風雪將他們困在2000公尺高的特拉基湖（現在稱為唐納湖畔。他們的食物逐漸消耗殆盡，12月中開始有人從營地出發尋找救援，但由於當時加州正經歷美墨戰爭缺乏人力，直到1847年2月中第一支搜救隊伍才終於抵達他們受困的湖畔營地，距離他們被困已將近四個月。87名成員中，最後只有48人活著抵達加州。

歷史學者將這一事件定位為西部移民史上最為慘痛的悲劇。[1]



派崔克·布林的日記第28頁，記載著他在1847年二月底的觀察。其中一段寫道：「莫菲太太昨日來此，說她正在考慮要開始吃掉繆特。我希望她應該還沒開始這麼做，這太讓人難過了。」



1859年，在內華達州洪堡河畔，移民團利用篷車建立起的營地

目次

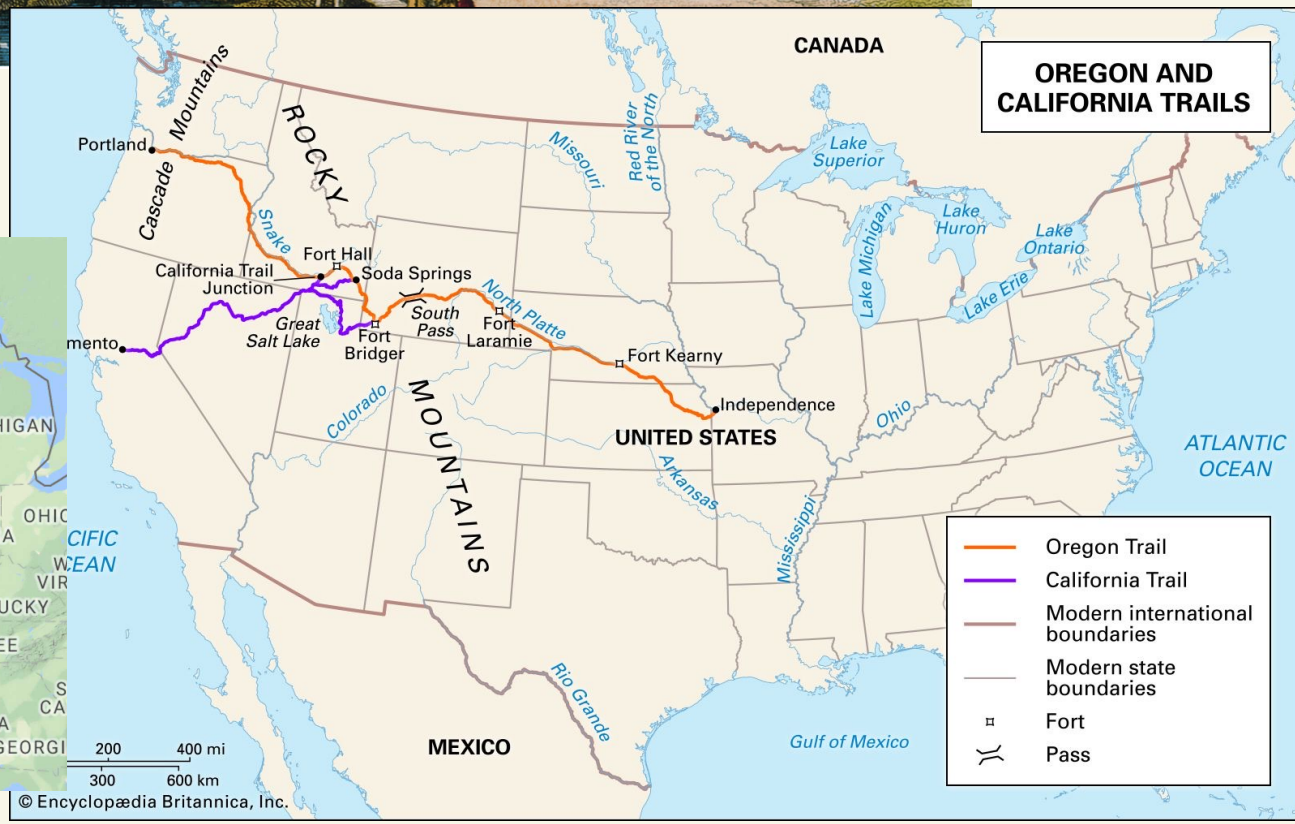
背景

1840年代，前往西部的移民者倍增。有的是像唐納大隊的成員派崔克·布林，嚮往能自由信仰公教的新天地而前往西部。也有許多人是被「昭昭天命」的想法所鼓舞，認為美國被賦予了向西擴張至橫跨北美洲大陸的天命，而積極前往西部定居。大多數馬車隊沿著從獨立城出發，到大陸分水嶺的俄勒岡小徑行途，每天行大約15英里（24公里）路，[2]來完成通常會費時4至6個月的路途。[3]這條路線通常沿著懷俄明的南部，在那裡有一塊山谷，可以使馬車隊更容易交換物資、補充資源。[4]自那裡起，馬車隊可以根據他們的目的地自由選擇路線。[5]

蘭斯福德·黑斯廷，一位早期移民，在1842年到了加利福尼亞州並看到了蕭條的現狀。為了鼓勵移民，他出版了 *俄勒岡、加利福尼亞移民指南* 。[6]他描述了一條可以直接穿過大盆地的近路，這條近路將帶領移民們通過沃薩奇山脈並穿過大鹽湖沙漠。黑斯廷並沒有穿越過他所設想的近路的任何


搜尋圖片內容


WASHINGTON
MONTANA
NORTH DAKOTA
MINNESOTA
SOUTH DAKOTA
WISCONSIN
MICHIGAN
OREGON
IDAHO
NEBRASKA
IOWA
Chicago
NEVADA
UTAH
COLORADO
KANSAS
MISSOURI
INDIANA
United States
CALIFORNIA
Las Vegas
Los Angeles
San Diego
ARIZONA
NEW MEXICO
OKLAHOMA
ARKANSAS
TENNESSEE
KENTUCKY
Dallas
MISSISSIPPI
ALABAMA
TEXAS
LOUISIANA
1
2
3
4
5
6
7
8


OREGON AND CALIFORNIA TRAILS
CANADA
Portland
Cascade Mountains
ROCKY MOUNTAINS
Snake
Missouri
Red River of the North
Lake Superior
Lake Michigan
Lake Huron
Lake Ontario
Lake Erie
Fort Hall
California Trail Junction
Soda Springs
Great Salt Lake
Fort Bridger
South Pass
Fort Laramie
North Platte
Fort Kearny
Independence
Colorado
Ohio
UNITED STATES
Arkansas
Mississippi
ATLANTIC OCEAN
Rio Grande
MEXICO
Gulf of Mexico
Oregon Trail
California Trail
Modern international boundaries
Modern state boundaries
Fort
Pass
200 400 mi
300 600 km
© Encyclopædia Britannica, Inc.

Example - Donner Party - Data

This data contains the ages and sexes of the adult (over 15 years) survivors and nonsurvivors of the Donner party.

	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Example - Donner Party - EDA

Status vs. Gender:

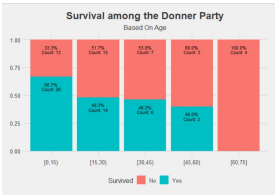
	Male	Female
Died	20	5
Survived	10	10

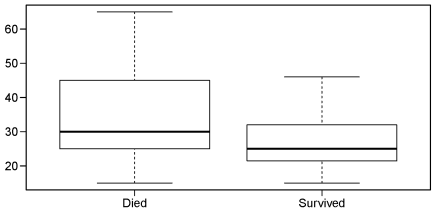
Die

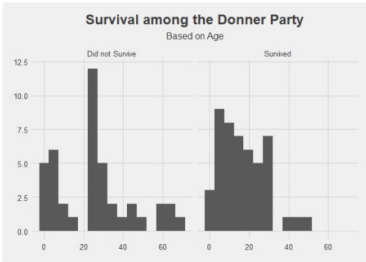
$\frac{P(E|M)}{P(E^c|M)} = \frac{2}{1}$ $\frac{P(E|F)}{P(E^c|F)} = \frac{1}{5}$

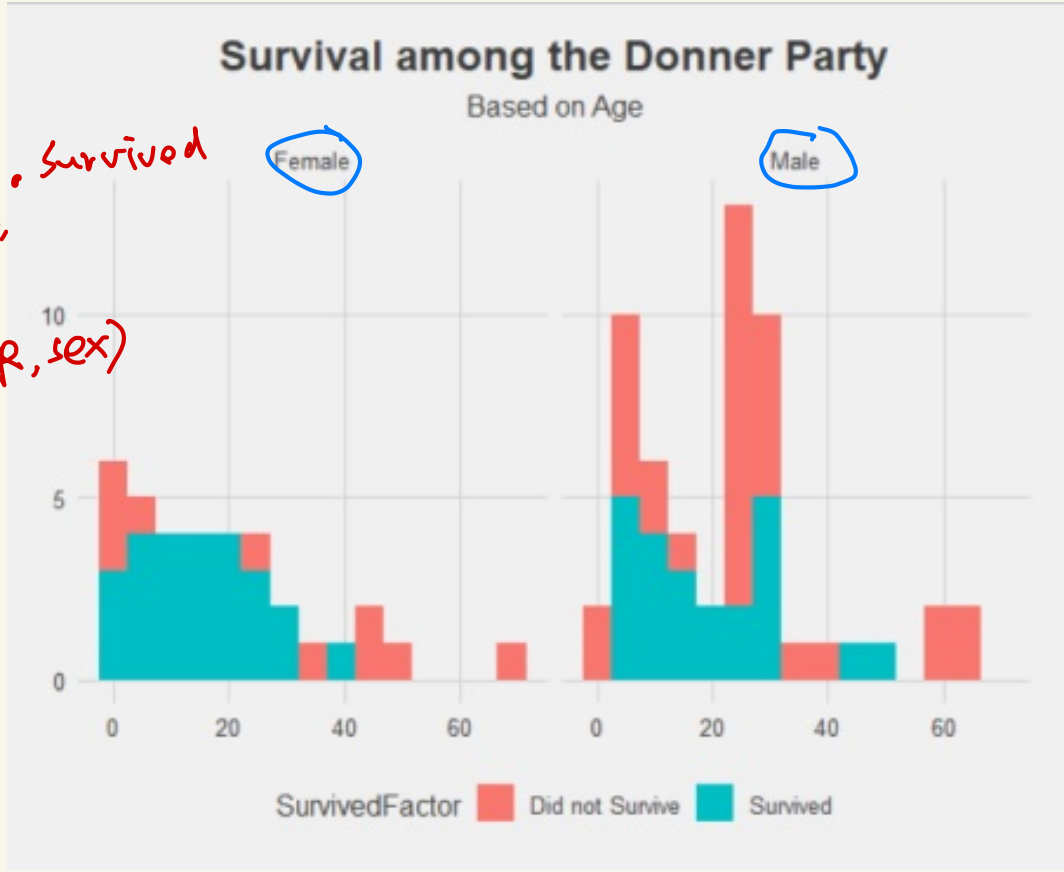
Odds of Mortality: Male 2 (2:1), Female 0.5 (1:2)

Status vs. Age:



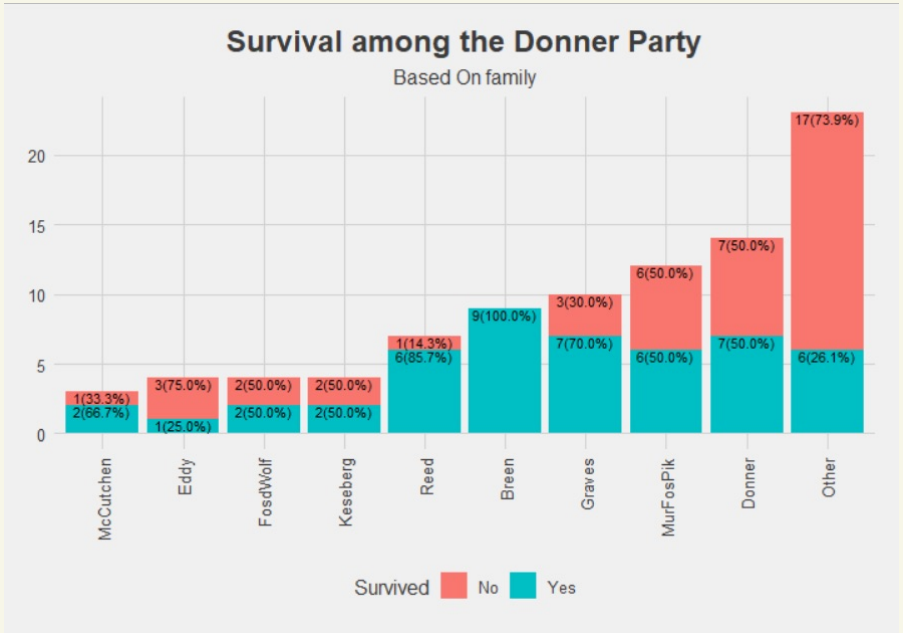






$Y = \begin{cases} 1, & \text{survived} \\ 0, & \end{cases}$

$P(Y=1 \mid age, sex)$



Example - Donner Party: Modeling Survival

It appears that both age and gender significantly influence survival outcomes. How can we develop a model to further explore these relationships?

$Y = \begin{cases} 1, & \text{survived} \\ 0, & \text{died} \end{cases}$

Simply coding the outcomes as 0 for Died and 1 for Survived doesn't sufficiently address the complexity of the problem. We need a more robust approach.

One effective method is to view survival in terms of a binomial distribution, where the probability of survival (success) versus non-survival (failure) can be modeled using a logistic function applied to a linear combination of predictors (age and gender).

Generalized linear models

It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called generalized linear models (GLMs). Logistic regression is just one example of this type of model.

All generalized linear models have the following three characteristics:

1. A probability distribution describing the outcome variable
2. A linear model
   - $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_r X_r$
3. A link function that relates the linear model to the parameter of the outcome distribution
   - $g(p) = \eta$ or $p = g^{-1}(\eta)$

Logistic Regression

Logistic Regression

$P(Y=1|x) = p(x)$, $P(Y=0|x) = 1-p(x)$

Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors.

We assume a binomial distribution produced the outcome variable and we therefore want to model p the probability of success for a given set of predictors.

To finish specifying the Logistic model we just need to establish a reasonable link function that connects η to p. There are a variety of options but the most commonly used is the logit function.

Logit function

$$logit(p) = \log\left(\frac{p}{1-p}\right), \text{ for } 0 \leq p \leq 1$$

$\ln\left(\frac{p(x)}{1-p(x)}\right) = A + \beta_1 x$

Properties of the Logit

The logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and ∞.

Inverse logit (logistic) function

$$g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$$

$$P(x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

The inverse logit function takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1.

This formulation also has some use when it comes to interpreting the model as logit can be interpreted as the log odds of a success, more on this later.

Logistic function

Article Talk

Language Watch Edit

For the recurrence relation, see Logistic map.

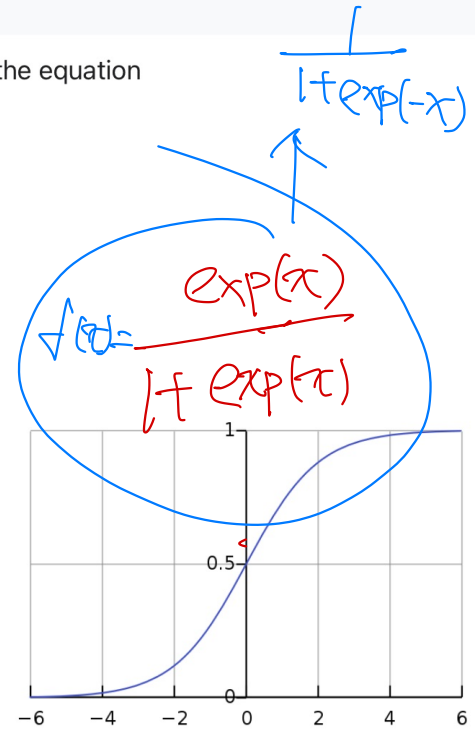
A **logistic function** or **logistic curve** is a common S-shaped curve (sigmoid curve) with the equation

$$f(x) = \frac{1}{1 + e^{-k(x-x_0)}}$$

where

x_0, the x value of the function's midpoint;
L, the supremum of the values of the function;
k, the logistic growth rate or steepness of the curve.[1]

For values of x in the domain of real numbers from $-\infty$ to $+\infty$, the S-curve shown on the right is obtained, with the graph of f approaching L as x approaches $+\infty$ and approaching zero as x approaches $-\infty$.



Standard logistic function where $L = 1, k = 1, x_0 = 0$

The logistic function finds applications in a range of fields, including biology (especially ecology), biomathematics, chemistry, demography, economics, geoscience, mathematical psychology, probability, sociology, political science, linguistics, statistics, and artificial neural networks. A generalization of the logistic function is the hyperbolastic function of type I.

The standard logistic function, where $L = 1, k = 1, x_0 = 0$, is sometimes simply called *the sigmoid*.[2] It is also sometimes called the *expit*, being the inverse of the logit.[3][4]

$P(Y=1|X) = p(x)$, $P(Y=0|X) = 1 - p(x)$

$$p(x) = \frac{1}{1 + \exp(-\beta_0 - \beta x)}$$

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

$$\text{odds}(Y=1|X) = e^{\beta_0 + \beta_1 x}$$

$$P(Y=0|X) = 1 - p(x) = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}$$

$$\ln(\text{odds}(Y=1|X)) = \beta_0 + \beta x$$

Logit

Article Talk

Language

Watch Edit

This article is about the binary logit function. For other types of logit, see discrete choice. For the basic regression technique that uses the logit function, see logistic regression. For standard magnitudes combined by multiplication, see logit (unit).

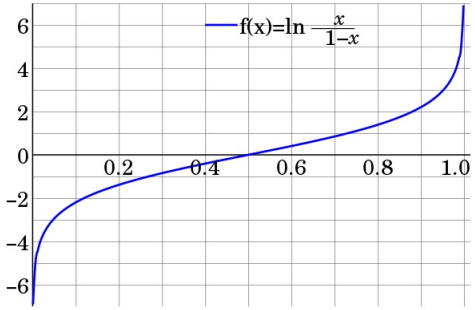
Not to be confused with log probability.

In statistics, the **logit** (/ˈloʊdʒɪt/ *LOH-jit*) function is the quantile function associated with the standard logistic distribution. It has many uses in data analysis and machine learning, especially in data transformations.

Mathematically, the logit is the inverse of the standard logistic function $\sigma(x) = 1/(1 + e^{-x})$, so the logit is defined as

$$\operatorname{logit} p = \sigma^{-1}(p) = \ln \frac{p}{1-p} \quad \text{for} \quad p \in (0, 1).$$

Because of this, the logit is also called the **log-odds** since it is equal to the logarithm of the odds $\frac{p}{1-p}$ where p is a probability. Thus, the logit is a type of function that maps probability values from $(0, 1)$ to real numbers in $(-\infty, +\infty)$,[1] akin to the probit function.



Plot of logit(x) in the domain of 0 to 1, where the base of the logarithm is e.

The logistic regression model

The three GLM criteria give us:

$$y_i \sim \text{Binom}(1, p_i)$$

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r$$

$$\text{logit}(p) = \eta$$

$\ln\left(\frac{p}{1-p}\right) = \eta = \beta' x$

$\frac{p}{1-p} = \exp(\eta) = \exp(\beta' x)$

$p = \frac{e^{\eta}}{1+e^{\eta}}$

From which we arrive at,

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_r x_{r,i})}{1 + \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_r x_{r,i})}$$

Example - Donner Party - Model

fit0 ← lm(y ~ x, data=..)
fit1 ← glm (y ~ x, data=..., family = bino)

In R we fit a GLM in the same was as a linear model except using glm instead of lm and we must also specify the type of GLM to fit using the family argument.

```
summary(glm(Status ~ Age, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.81852    0.99937   1.820   0.0688 .
## Age         -0.06647    0.03222  -2.063   0.0391 *
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 56.291  on 43  degrees of freedom
## AIC: 60.291
##
## Number of Fisher Scoring iterations: 4
```

Status = { 1, survive.
0. die

Example - Donner Party - Prediction

LR
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$P(Y=1 \mid \text{Age})$

$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$

Odds / Probability of survival for a newborn (Age=0):

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

$$\frac{p}{1-p} = \exp(1.8185) = 6.16 \quad \text{odds}$$

$$p = 6.16/7.16 = 0.86$$

Example - Donner Party - Prediction (cont.)

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a 25 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 25$$

$$\frac{p}{1-p} = \exp(0.156) = 1.17 \quad \text{odds}$$

$$p = 1.17/2.17 = 0.539$$

odds > 1
p > 0.5

Odds / Probability of survival for a 50 year old:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 50$$

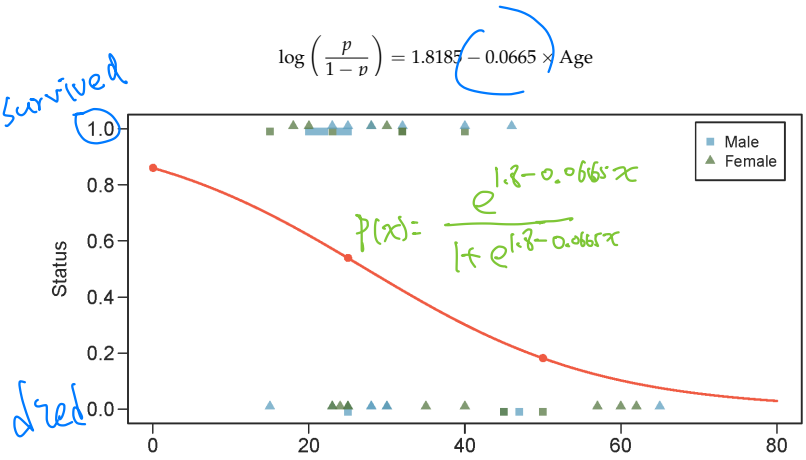
$$\frac{p}{1-p} = \exp(-1.5065) = 0.222 \quad \text{odds}$$

$$p = 0.222/1.222 = 0.181$$

odds < 1
p < 0.5

Example - Donner Party - Prediction (cont.)

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$



Example - Donner Party - Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

$\ln\frac{P(x)}{1-P(x)} = \hat{\beta}_0 - 0.0665x$

$\text{odds}(x)$

$\ln(\text{odds}(x+1)) - \ln(\text{odds}(x)) = -0.0665$

$\ln\left[\frac{\text{odds}(x+1)}{\text{odds}(x)}\right] = -0.0665$

OR

Interpretation:

- Intercept: The log odds of survival for a party member hypothetically aged 0. This base value allows for calculation of odds and probability with further computation.
- Slope (Age): The coefficient for age, -0.0665, represents the change in the log odds of survival for each additional year of age. This can be converted to an odds ratio (OR):

$$\text{OR} = e^{-0.0665} \approx 0.935$$

This means that each additional year of age is associated with a 6.5% decrease in the odds of survival (since $1 - 0.935 = 0.065$), emphasizing the negative impact of increasing age on survival prospects.

Understanding the Odds Ratio

Definition: Odds Ratio (OR) is a statistic that quantifies the strength and direction of the association between two binary variables. It compares the odds of an event occurring in one group to the odds of it occurring in another group.

$$\text{Odds Ratio (OR)} = \frac{\text{Odds of event in the exposed group}}{\text{Odds of event in the unexposed group}} = \frac{\frac{P(E=1|X=1)}{P(E=0|X=1)}}{\frac{P(E=1|X=0)}{P(E=0|X=0)}}$$

$$\frac{\text{Odds in age } (x+1)}{\text{odds in age } (x)} = e^{-0.0665}$$

$\Rightarrow$ indep. of X

- **OR**$= 1$: No association between exposure and outcome.
- **OR**> 1: Positive association; greater odds of the event occurring with the exposure.
- **OR**< 1: Negative association; lower odds of the event occurring with the exposure.

Application in Logistic Regression: In logistic regression, the exponentiated coefficient (e^{β}) of a predictor variable gives the odds ratio. This measures how the odds of the outcome change with a one-unit increase in the predictor, holding all else constant.

Example - Donner Party - Interpretation - Slope

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.8185 - 0.0665(x+1)$$

$$= 1.8185 - 0.0665x - 0.0665$$

$$\log\left(\frac{p_0}{1-p_0}\right) = 1.8185 - 0.0665x$$

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_0}{1-p_0}\right) = -0.0665$$

$$\log\left(\frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0}\right) = -0.0665$$

$$\frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0} = \exp(-0.0665) = 0.94$$

odds ratio

Example - Donner Party - Age and Gender

multiple covariates

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

$\ln \frac{p_i}{1-p_i} = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i$

$Y_i \sim Ber(p_i)$

$Sex = \begin{cases} 1 & \text{Female} \\ 0 & \text{Male} \end{cases}$

$P(Y_i = 1) = p_i$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

AIC = 60

Gender slope: When the other predictors are held constant this is the log odds ratio between the given level (Female) and the reference level (Male).

Example - Donner Party - Gender Models

Just like MLR we can plug in gender to arrive at two status vs age models for men and women respectively.

General model:

$$\log\left(\frac{p_1}{1-p_1}\right) = 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times \text{Sex}$$

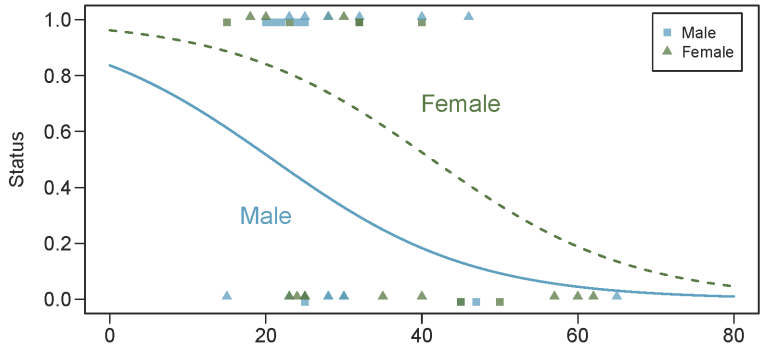
Male model:

$$\begin{aligned}\log\left(\frac{p(\text{x})}{1-p(\text{x})}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times 0 \\ &= 1.63312 + -0.07820 \times \text{Age}\end{aligned}$$

Female model:

$$\begin{aligned}\log\left(\frac{p(\text{x})}{1-p(\text{x})}\right) &= 1.63312 + -0.07820 \times \text{Age} + 1.59729 \times 1 \\ &= 3.23041 + -0.07820 \times \text{Age}\end{aligned}$$

Example - Donner Party - Gender Models (cont.)



Hypothesis test for the whole model

$P(x,g) = P(Y=1 \mid x, g)$

$\ln\left(\frac{P(x,g)}{1-P(x,g)}\right) = \beta_0 + \beta_1 x + \beta_2 g$

```
summary(glm(Status ~ Age + Sex, data=donner, family=binomial))

## Call:
## glm(formula = Status ~ Age + Sex, family = binomial, data = donner)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## SexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Fix Age = x

$\ln\left(\frac{\text{Odds}(x, g=1)}{\text{Odds}(x, g=0)}\right) = \beta_2$

$\Rightarrow$ OR for gender $= e^{\beta_2}$

$e^{1.59} \doteq 4.9$

$\text{odds}(x, g=1) = 4.9 \cdot \text{odds}(x, g=0)$

Note: The model output does not include any F-statistic.

Hypothesis tests for a coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

We are however still able to perform inference on individual coefficients, the basic setup is exactly the same as what we've seen before except we use a Z test.

Note: The only tricky bit, which is way beyond the scope of this course, is how the standard error is calculated.

Testing for the slope of Age

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

$$H_0 : \beta_{age} = 0$$
$$H_A : \beta_{age} \neq 0$$

$$Z = \frac{\hat{\beta}_{age} - \beta_{age}}{SE_{age}} = \frac{-0.0782 - 0}{0.0373} = -2.10$$

$$\text{p-value} = P(|Z| > 2.10) = P(Z > 2.10) + P(Z < -2.10)$$
$$= 2 \times 0.0178 = 0.0359$$

$N(0,1)$

Confidence interval for age slope coefficient

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.6331	1.1102	1.47	0.1413
Age	-0.0782	0.0373	-2.10	0.0359
SexFemale	1.5973	0.7555	2.11	0.0345

Remember, the interpretation for a slope is the change in log odds ratio per unit change in the predictor.

95% CI for β_1

Log odds ratio: β_1

$$CI = PE \pm CV \times SE = -0.0782 \pm 1.96 \times 0.0373 = (-0.151, -0.005)$$

confidence interval; PE: $\hat{\beta}_1$; CV: $z_{97.5}$; SE: S.E.

Odds ratio: e^{β_1}

$$\exp(CI) = (e^{-0.151}, e^{-0.005}) = (0.859, 0.995)$$

$\ln(\text{Odds}(x,g)) = \beta_0 + \beta_1 x + \beta_2 g + \beta_3 xg$ Let $OR(x) = \dfrac{\text{odds}(age=x, g=1)}{\text{odds}(age=x, g=0)}$

glm(Status ~ Age*Sex, data=donner, family=binomial)

$$\ln\left(\frac{\text{odds}(x,g=1)}{\text{odds}(x,g=0)}\right) = \beta_2 + \beta_3 x$$

$$OR(x) = e^{\beta_2 + \beta_3 x}$$

depends on x

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.24638    3.20517   2.261   0.0238 *
age         -0.19407    0.08742  -2.220   0.0264 *
sex         -6.92805    3.39887  -2.038   0.0415 *
age:sex      0.16160    0.09426   1.714   0.0865 .
```

For females:

age

$$\hat{\pi} = \frac{\exp(7.2450 - 0.1940 \times 24 - 6.9267 \times 0 + 0.1616 \times 0)}{1 + \exp(7.2450 - 0.1940 \times 24 - 6.9267 \times 0 + 0.1616 \times 0)} = 0.930$$

The odds of surviving are multiplied by a factor of $\exp(-0.194) = 0.824$ per additional year of age.

For males:

age >4

$$\hat{\pi} = \frac{\exp(7.2450 - 0.1940 \times 24 - 6.9267 \times 1 + 0.1616 \times (24 \times 1))}{1 + \exp(7.2450 - 0.1940 \times 24 - 6.9267 \times 1 + 0.1616 \times (24 \times 1))} = \frac{\exp(0.3183 - 0.0324 \times 24)}{1 + \exp(0.3183 - 0.0324 \times 24)} = 0.387$$

The odds of surviving are multiplied by a factor of $\exp(-0.0324) = .968$ per additional year of age. Note that the odds of survival decline more rapidly for females than for males.

$\ln(\text{odds}(x, g=1)) = \beta_0 + \beta_1 x + \beta_2 + \beta_3 x$

$\ln(\text{odds}(x, g=0)) = \beta_0 + \beta_1 x$

Additional Example

Example - Birdkeeping and Lung Cancer

A health survey conducted from 1972 to 1981 in The Hague, Netherlands, revealed an association between keeping pet birds and an increased risk of lung cancer. In response, researchers initiated a case-control study in 1985 at four hospitals within The Hague (population: 450,000).

Study Details:

- **Cases:** 49 lung cancer patients, registered with a general practice, aged 65 or younger, and residents of the city since at least 1965.
- **Controls:** 98 residents selected to match the general age structure of the cases.

This study aimed to explore birdkeeping as a potential risk factor for lung cancer among the population.

Source: Ramsey, F.L. and Schafer, D.W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis* (2nd edition).

Example - Birdkeeping and Lung Cancer - Data

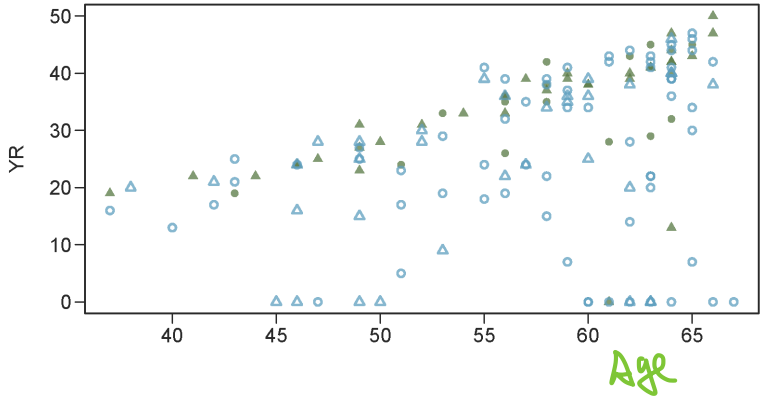
	LC	FM	SS	BK	AG	YR	CD
1	LungCancer	Male	Low	Bird	37.00	19.00	12.00
2	LungCancer	Male	Low	Bird	41.00	22.00	15.00
3	LungCancer	Male	High	NoBird	43.00	19.00	15.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
147	NoCancer	Female	Low	NoBird	65.00	7.00	2.00

$LC = \begin{cases} 1, & \text{Cancer} \\ 0, & \text{no} \end{cases}$

LC	Whether subject has lung cancer
FM	Sex of subject
SS	Socioeconomic status
BK	Indicator for birdkeeping
AG	Age of subject (years)
YR	Years of smoking prior to diagnosis or examination
CD	Average rate of smoking (cigarettes per day)

Note:NoCancer is the reference response (0 or failure), LungCancer is the non-reference response (1 or success) - this matters for interpretation.

Example - Birdkeeping and Lung Cancer - EDA



Example - Birdkeeping and Lung Cancer - Model

```
summary(glm(LC ~ FM + SS + BK + AG + YR + CD, data=bird, family=binomial))

## Call:
## glm(formula = LC ~ FM + SS + BK + AG + YR + CD, family = binomial,
##     data = bird)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.93736    1.80425  -1.074 0.282924
## FMFemale     0.56127    0.53116   1.057 0.290653
## SSHigh       0.10545    0.46885   0.225 0.822050
## BKBird       1.36259    0.41128   3.313 0.000923 ***
## AG          -0.03976    0.03548  -1.120 0.262503
## YR           0.07287    0.02649   2.751 0.005940 **
## CD           0.02602    0.02552   1.019 0.308055
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.14  on 146  degrees of freedom
## Residual deviance: 154.20  on 140  degrees of freedom
## AIC: 168.2
##
## Number of Fisher Scoring iterations: 5
```

Example - Birdkeeping and Lung Cancer - Interpretation

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9374	1.8043	-1.07	0.2829
FMFemale	0.5613	0.5312	1.06	0.2907
SSHigh	0.1054	0.4688	0.22	0.8221
BKBird	1.3626	0.4113	3.31	0.0009
AG	-0.0398	0.0355	-1.12	0.2625
YR	0.0729	0.0265	2.75	0.0059
CD	0.0260	0.0255	1.02	0.3081

Keeping all other predictors constant then,

- The odds ratio of getting lung cancer for bird keepers vs non-bird keepers is $\exp(1.3626) = 3.91$.
- The odds ratio of getting lung cancer for an additional year of smoking is $\exp(0.0729) = 1.08$.

What do the numbers not mean ...

The most common mistake made when interpreting logistic regression is to treat an odds ratio as a ratio of probabilities.

Bird keepers are *not* 4x more likely to develop lung cancer than non-bird keepers.

This is the difference between relative risk and an odds ratio.

$$RR = \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})}$$

$$OR = \frac{P(\text{disease}|\text{exposed})/[1 - P(\text{disease}|\text{exposed})]}{P(\text{disease}|\text{unexposed})/[1 - P(\text{disease}|\text{unexposed})]}$$

Back to the birds

What is probability of lung cancer in a bird keeper if we knew that $P(\text{lung cancer}|\text{no birds}) = 0.05$?

$$OR = \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{P(\text{lung cancer}|\text{no birds})/[1 - P(\text{lung cancer}|\text{no birds})]}$$

$$= \frac{P(\text{lung cancer}|\text{birds})/[1 - P(\text{lung cancer}|\text{birds})]}{0.05/[1 - 0.05]} = 3.91$$

$$P(\text{lung cancer}|\text{birds}) = \frac{3.91 \times \frac{0.05}{0.95}}{1 + 3.91 \times \frac{0.05}{0.95}} = 0.171$$

$$RR = P(\text{lung cancer}|\text{birds})/P(\text{lung cancer}|\text{no birds}) = 0.171/0.05 = 3.41$$

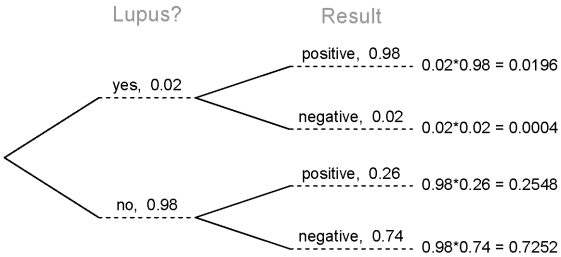
Sensitivity and Specificity

(An old) Example - *House*

If you are familiar with the TV show *House* on Fox, you might recall Dr. House's frequent remark: "It's never lupus."

- **What is Lupus?**
  - Lupus is an autoimmune disease where antibodies, instead of protecting against infections, mistakenly target the body's own proteins as foreign invaders.
  - This abnormal immune response can lead to increased blood clotting risks.
  - Approximately 2% of the population suffers from lupus.
- **Diagnostic Accuracy:**
  - If a person has lupus, the test is 98% accurate.
  - If a person does not have lupus, the test is 74% accurate.
- **Discussion Point:**
  - Considering the test's accuracy, is Dr. House correct in his skepticism even if a test result is positive for lupus?

(An old) Example - *House*



$$P(\text{Lupus}|+) = \frac{P(+,\text{Lupus})}{P(+,\text{Lupus}) + P(+,\text{No Lupus})}$$
$$= \frac{0.0196}{0.0196 + 0.2548} = 0.0714$$

Testing for Lupus

Diagnosing lupus involves a complex array of multiple tests, reflecting the multifaceted nature of the disease.

- **Common Tests for Lupus:**
  - **Complete Blood Count (CBC):** Assesses overall health and detects disorders like anemia, infection, and other diseases.
  - **Erythrocyte Sedimentation Rate (ESR):** Measures the rate at which red blood cells sediment in a period of one hour, indicating inflammation.
  - **Kidney and Liver Assessment:** Evaluates the function of these organs which can be affected by lupus.
  - **Urinalysis:** Tests for protein or red blood cells in the urine, which indicates kidney damage.
  - **Antinuclear Antibody (ANA) Test:** Detects antibodies that often are present in individuals with autoimmune diseases like lupus.

Testing for Lupus: Binary Decision Making

Diagnosing lupus can be viewed as a binary decision (lupus or no lupus) that requires the integration of multiple test results.

- **Binary Decision:**
  - The decision involves considering lupus or not based on a range of explanatory variables derived from various tests.
- **Importance of Sensitivity and Specificity:**
  - Sensitivity indicates the test's ability to correctly identify those with the disease (true positive rate).
  - Specificity refers to the test's ability to correctly identify those without the disease (true negative rate).
  - These metrics help interpret what a positive or negative test result actually means in the context of diagnosing lupus.

Sensitivity and Specificity

Sensitivity - measures a tests ability to identify positive results.

$$P(\text{Test } + \mid \text{Conditon } +) = P(+|\text{lupus}) = 0.98$$

Specificity - measures a tests ability to identify negative results.

$$P(\text{Test } - \mid \text{Condition } -) = P(-|\text{no lupus}) = 0.74$$

It is illustrative to think about the extreme cases - what is the sensitivity and specificity of a test that always returns a positive result? What about a test that always returns a negative result?

Sensitivity and Specificity (cont.)

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I error)
Test Negative	False Negative (Type II error)	True Negative

$$
\begin{aligned}
Sensitivity &= P(\text{Test } + \mid \text{Condition } +) = TP/(TP+FN) \\
Specificity &= P(\text{Test } - \mid \text{Condition } -) = TN/(FP+TN) \\
Falsenegativerate(\beta) &= P(\text{Test } - \mid \text{Condition } +) = FN/(TP+FN) \\
Falsepositiverate(\alpha) &= P(\text{Test } + \mid \text{Condition } -) = FP/(FP+TN)
\end{aligned}
$$

So What?

Understanding test sensitivity, specificity, and disease incidence is crucial for accurate medical decision-making.

- **Key Measures:**
  - Sensitivity and specificity help calculate probabilities like $P(\text{lupus}|+)$.
- **Using This Information:**
  - How do we apply these insights to improve diagnostic decisions?

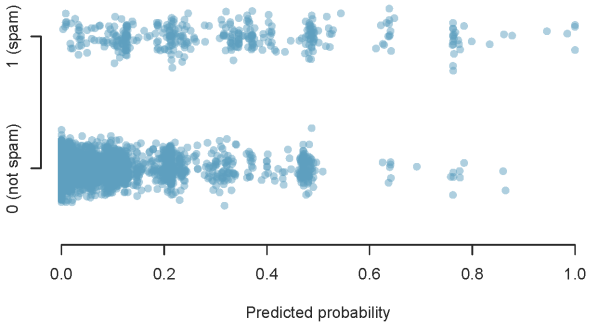
ROC curves

Identifying Spam Messages

Our analysis involved using logistic regression models to determine the likelihood of emails being spam based on various predictors.

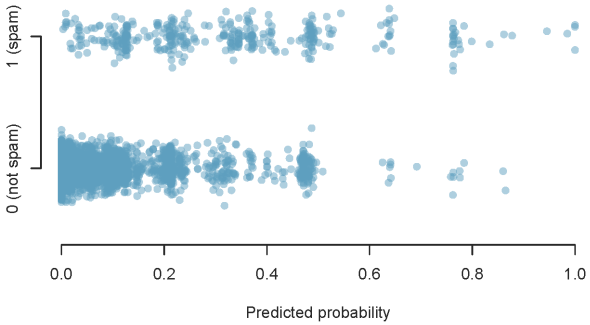
- **Purpose of Models:**
  - Assess influence of different predictors on spam classification.
  - Assign probabilities to incoming messages for real-time filtering.
- **Beyond Probability Assignment:**
  - Effective spam filtering requires decision-making based on assigned probabilities.
- **Decision Rule:**
  - Implement a simple threshold probability.
  - Emails exceeding this threshold are flagged as spam.

Picking a threshold



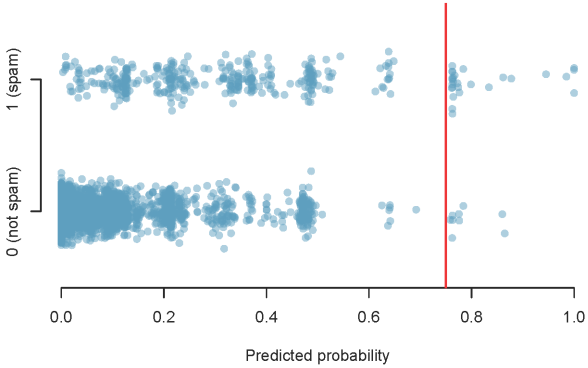
Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



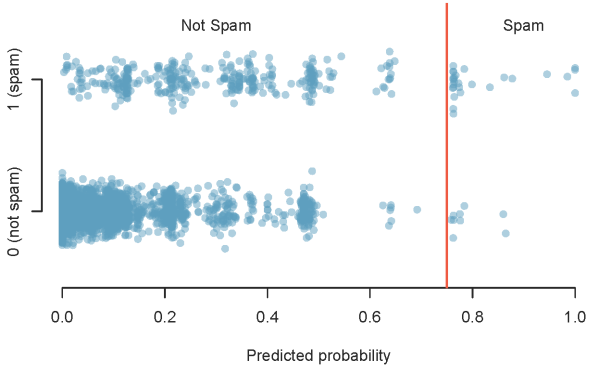
Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



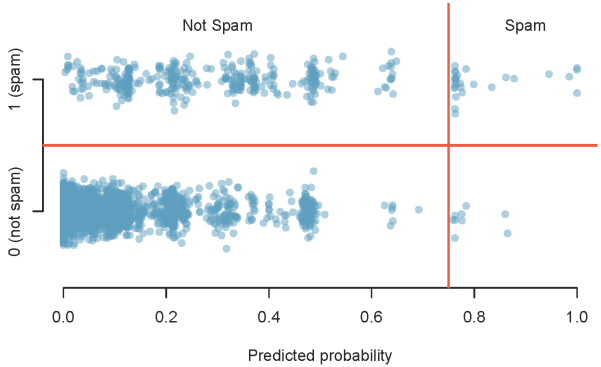
Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



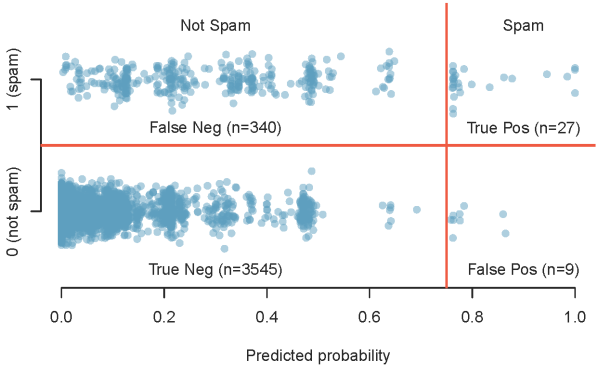
Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



Lets see what happens if we pick our threshold to be 0.75.

Consequences of picking a threshold

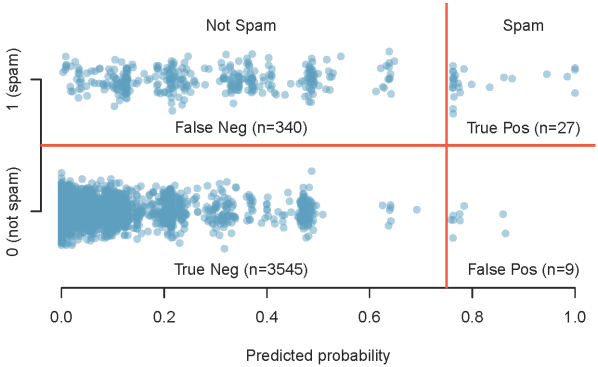
For our data set picking a threshold of 0.75 gives us the following results:

$$FN = 340 \qquad TP = 27$$
$$TN = 3545 \qquad FP = 9$$

What are the sensitivity and specificity for this particular decision rule?

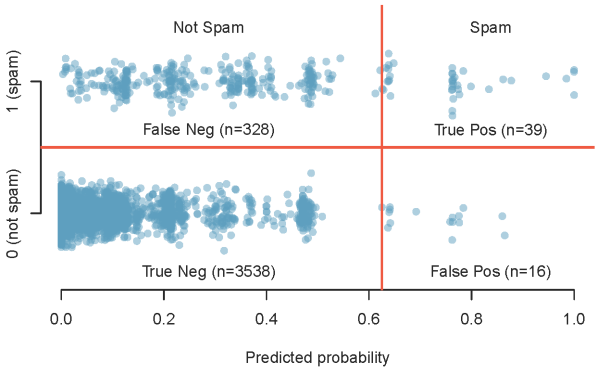
$$\text{Sensitivity} = TP/(TP + FN) = 27/(27 + 340) = 0.073$$
$$\text{Specificity} = TN/(FP + TN) = 3545/(9 + 3545) = 0.997$$

Trying other thresholds



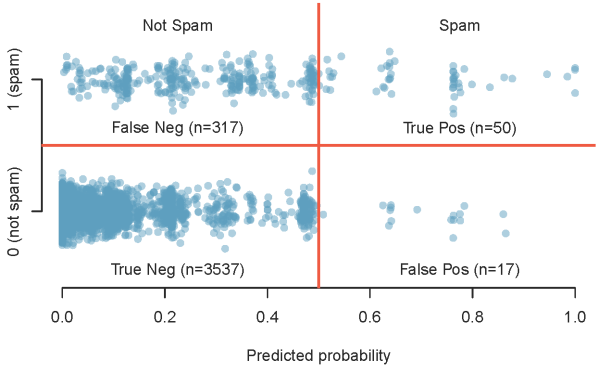
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106			
Specificity	0.997	0.995			

Trying other thresholds



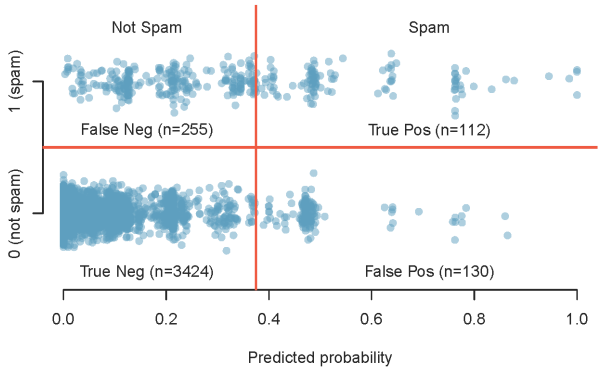
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106			
Specificity	0.997	0.995			

Trying other thresholds



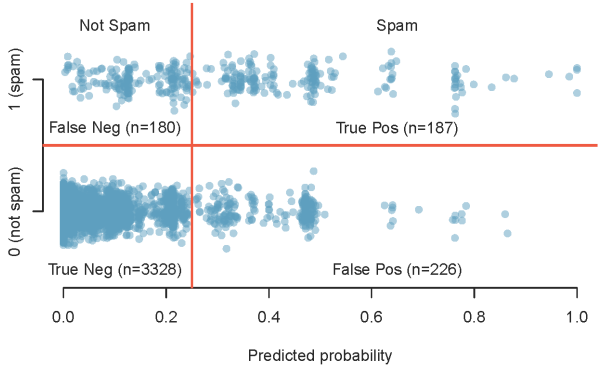
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136		
Specificity	0.997	0.995	0.995		

Trying other thresholds



Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	
Specificity	0.997	0.995	0.995	0.963	

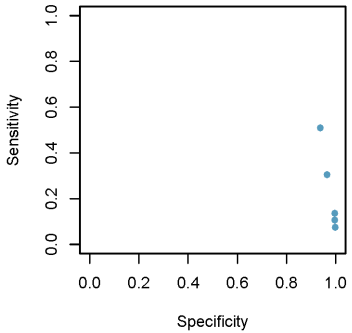
Trying other thresholds



Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936

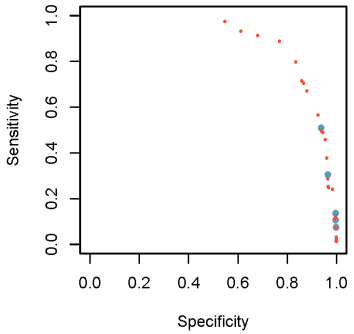
Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936



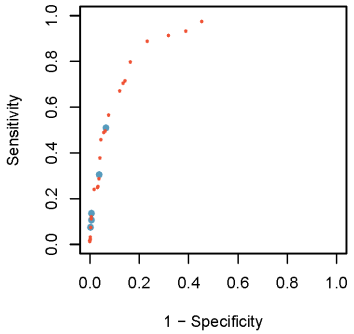
Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936

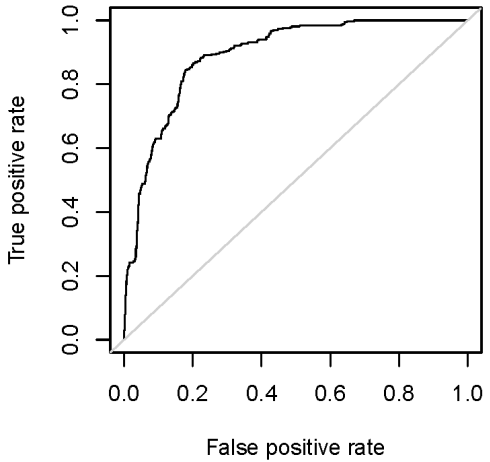


Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936



Receiver operating characteristic (ROC) curve



Receiver operating characteristic (ROC) curve (cont.)

Why do we care about ROC curves?

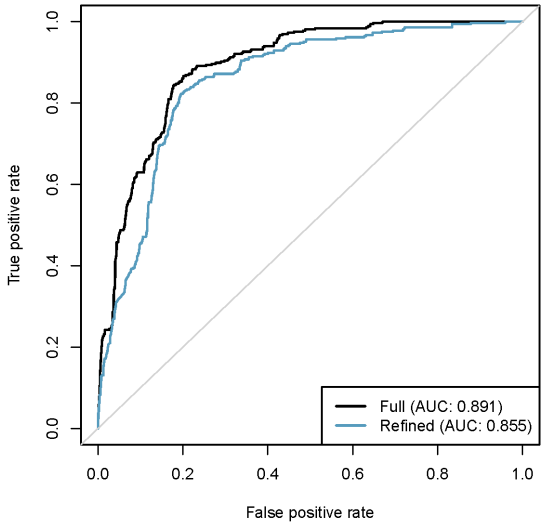
- Shows the trade off in sensitivity and specificity for all possible thresholds.
- Straight forward to compare performance vs. chance.
- Can use the area under the curve (AUC) as an assessment of the predictive ability of a model.

Refining the Spam model

```
g_refined = glm(spam ~ to_multiple+cc+image+attach+winner
                     +password+line_breaks+format+re_subj
                     +urgent_subj+exclaim_mess,
                data=email, family=binomial)
summary(g_refined)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.7594	0.1177	-14.94	0.0000
to_multipleyes	-2.7368	0.3156	-8.67	0.0000
ccyes	-0.5358	0.3143	-1.71	0.0882
imageyes	-1.8585	0.7701	-2.41	0.0158
attachyes	1.2002	0.2391	5.02	0.0000
winneryes	2.0433	0.3528	5.79	0.0000
passwordyes	-1.5618	0.5354	-2.92	0.0035
line_breaks	-0.0031	0.0005	-6.33	0.0000
formatPlain	1.0130	0.1380	7.34	0.0000
re_subjyes	-2.9935	0.3778	-7.92	0.0000
urgent_subjyes	3.8830	1.0054	3.86	0.0001
exclaim_mess	0.0093	0.0016	5.71	0.0000

Comparing models



Utility Functions

Utility Functions

There are many other reasonable quantitative approaches we can use to decide on what is the "best" threshold.

If you've taken an economics course you have probably heard of the idea of utility functions, we can assign costs and benefits to each of the possible outcomes and use those to calculate a utility for each circumstance.

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	
False Positive	
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	
False Positive	
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	-50
False Negative	

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	-50
False Negative	-5

Utility function for our spam filter

To write down a utility function for a spam filter we need to consider the costs / benefits of each out.

Outcome	Utility
True Positive	1
True Negative	1
False Positive	-50
False Negative	-5

$$U(p) = TP(p) + TN(p) - 50 \times FP(p) - 5 \times FN(p)$$

Utility for the 0.75 threshold

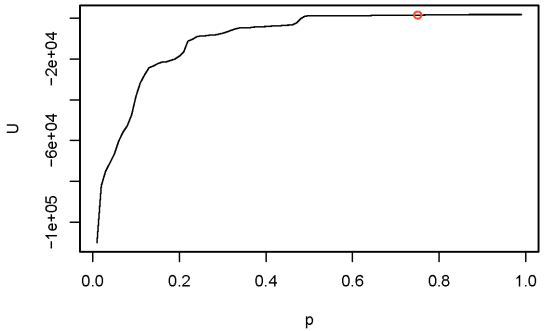
For the email data set picking a threshold of 0.75 gives us the following results:

$$FN = 340 \qquad TP = 27$$
$$TN = 3545 \qquad FP = 9$$

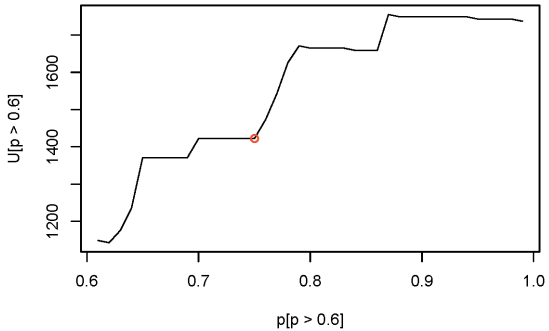
$$\begin{aligned} U(p) &= TP(p) + TN(p) - 50 \times FP(p) - 5 \times FN(p) \\ &= 27 + 3545 - 50 \times 9 - 5 \times 340 = 1422 \end{aligned}$$

Not useful by itself, but allows us to compare with other thresholds.

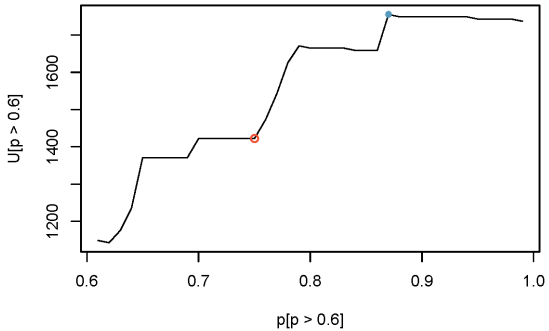
Utility curve



Utility curve (zoom)



Utility curve (zoom)



Maximum Utility