Logistic Regression

& R Programming

Simple Linear Regression

Modeling the relationship between two variables is an important task in science, business and everyday life.

The simplest model is the simple linear regression model:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

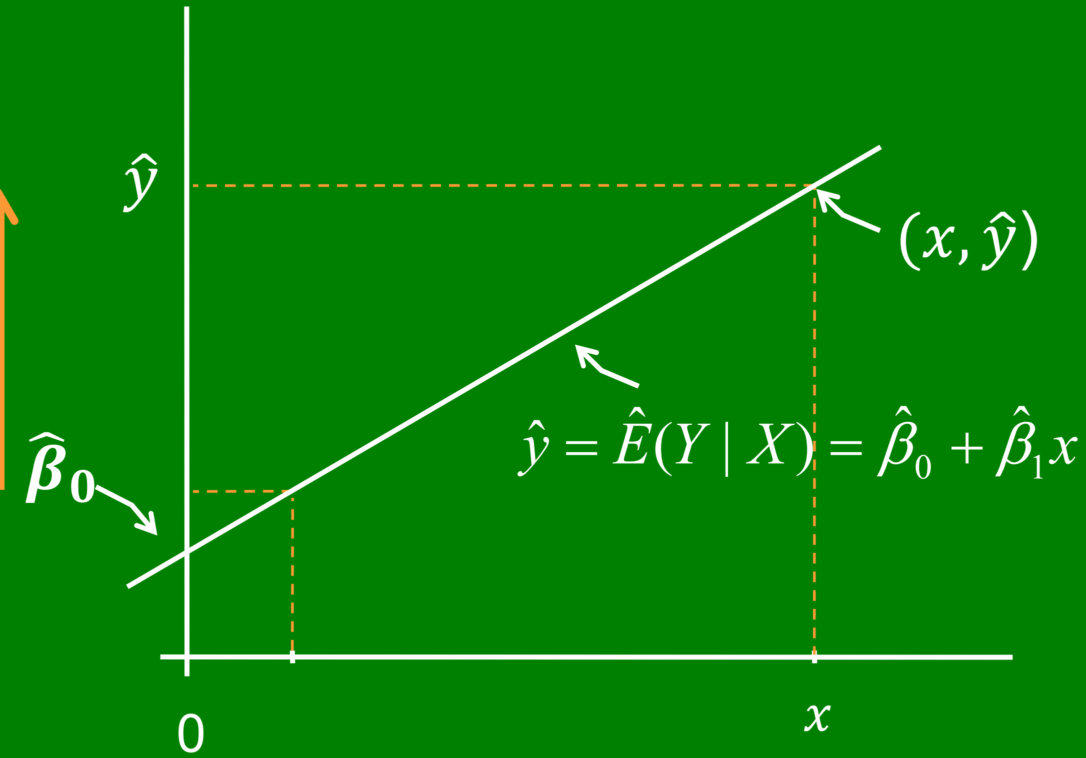
Here ε is a random error with mean zero.

Taking expectation at both sides of the equation, we see the simple linear regression models the mean of the response Y as a linear function of the predictor x:

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

$E(\varepsilon) = 0$ *(This link is called the identity link)*

Simple Linear Regression
The estimated regression equation



$\hat{\beta}_0$ = Estimated Intercept

= $\hat{y}$-value at $x = 0$

Interpretable only if x = 0 is a value of particular interest.

$\hat{\beta}_1$ = Estimated Slope

= Change in $\hat{y}$ for every unit increase in x

= estimated change in the mean of Y for a 1 unit change in X.

Always interpretable.

Multiple Linear Regression

We model the mean of a numeric response Y as a linear combination of p predictors or some functions of these predictors, i.e.

$$E(Y|\mathbf{X}) = \underbrace{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}_{\text{Here the terms in the model are the predictors}}$$

$$E(Y|\mathbf{X}) = \beta_o + \underbrace{\beta_1 f_1(\mathbf{X}) + \beta_2 f_2(\mathbf{X}) + \ldots + \beta_k f_k(\mathbf{X})}_{\text{Here the terms in the model are k different functions of the p predictors}}$$

Multiple Linear Regression

For the classic multiple regression model

$E(Y|\mathbf{X}) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

the regression coefficient (β_i) represents the estimated change in the mean of the response Y associated with a unit change in X_i while the other predictors are held constant.

The multiple linear regression model is called the general linear model when we have at least one categorical predictor in the model.

Generalized Linear Models

- Family of regression models
- Response (Y) | Model Type

Response (Y)	Model Type
– Continuous	General Linear Model
– Counts	Poisson regression
– Survival time	Cox regression model
– Binary	Logistic regression model

- Uses
  - Control for potentially confounding factors
  - Model building , risk prediction

Logistic Regression

- Models relationship between A dichotomous categorical response variable *Y*

  e.g. Success/Failure, Diseased/ Normal, Survived/Died, green eyes/not green eyes, vote for candidate A/do not vote for candidate A, etc…

and

- A set of predictor variables X_i:
  - dichotomous (yes/no, smoker/nonsmoker,…)
  - other categorical (social class, race, ...)
  - continuous (age, weight, gestational age, ...)

Categorical Response Variables

Whether or not a person smokes

Binary Response

$$Y = \begin{cases} \text{Non-smoker} \\ \text{Smoker} \end{cases}$$

Success of a medical treatment

$$Y = \begin{cases} \text{Survives} \\ \text{Dies} \end{cases}$$

Opinion poll responses

Ordinal Response

$$Y = \begin{cases} \text{Agree} \\ \text{Neutral} \\ \text{Disagree} \end{cases}$$

Example: Height predicts Gender

Y = Gender (0=Male 1=Female)

X = Height (Hgt, in inches)

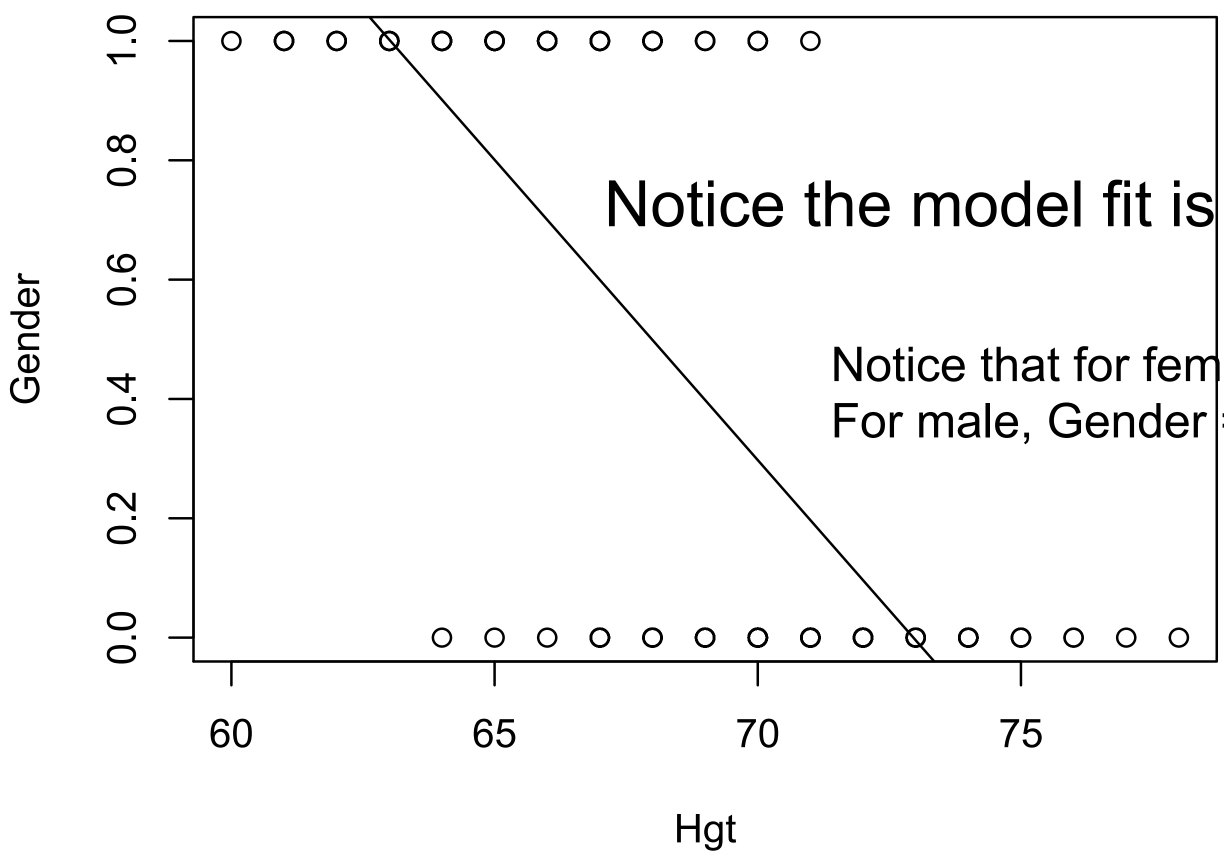
First we try the simple linear regression model:

```
> regmodel=lm(Gender~Hgt,data=Pulse)
> summary(regmodel)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.343647   0.397563   18.47   <2e-16 ***
Hgt         -0.100658   0.005817  -17.30   <2e-16 ***
```

This simple linear regression model does not fit the data well. In other words, linking the mean of the response variable to the predictor directly using the identity link function does not seem to be a good choice here when the response variable is binary.

We will have to use a different link function.



Notice the model fit is terrible.

Notice that for female: Gender =1;
For male, Gender = 0.

π = the Population Proportion of "Success"

In linear regression the model predicts the *mean* Y for a linear combination of predictors.

What's the mean of a 0/1 binary (indicator) variable?

$$\bar{y} = \frac{\sum y_i}{n} = \frac{\#\text{ of 1's}}{\#\text{ of trials}} = \text{Proportion of "success"}$$

= Sample Proportion: π_hat ($\hat{\pi}$)

* Goal of logistic regression: Predict the "true" proportion of success, π (sometimes we use p), at any value of the predictor x.

* For a binary response Y (0,1 valued),

P(Y=1) = π; E(Y) = μ =1*π+0*(1-π) = π

(Binary) Logistic Regression Model

Y = Binary response

X = Quantitative predictor

π = proportion of 1's (yes,success) at any X

When we fit the binary response to the predictor using the simple linear regression (with the identity link):

$$\pi(x) = \mu(x) = \beta_0 + \beta_1 * x$$

This is not reasonable

Because the left-side is in (0,1) and the right can be as much as $(-\infty, +\infty)$: different scales!

Binary Logistic Regression Model

Y = Binary response | X = Quantitative predictor

π = proportion of 1's (yes,success) at any X

Equivalent forms of the logistic regression model:

Logit form

$$ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x$$

(Note: some use log but it means ln)

Probability form

$$\pi = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

This is always the natural log with base e (aka "ln")

Ln[π/(1-π)] has the range of (-∞, +∞):

The same scales on both sides of the equation 

Note: For the logistic regression, we model the group with the same parameters β_0, β_1. However, each subject has its own predictor x.

How to fit the model?

OLS (ordinary least squares)? – questionable

Maximum likelihood estimators (MLE) – this is what we use to fit the model because each Y|x ~ Bernoulli(π)

PDF of Y|x: f(y|x) = $\pi^y(1-\pi)^{1-y}$

For a random sample of: $(x_i, y_i), i = 1, \dots, n$

Its likelihood function: $L = f(y_1, \dots, y_n) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}$

For the logistic regression we have: $\pi(x_i) = E(Y_i|x_i)$

$$ln \frac{\pi(x_i)}{1-\pi(x_i)} = \beta_0 + \beta_1 x_i, i = 1, \dots, n$$

The likelihood function: $L = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$

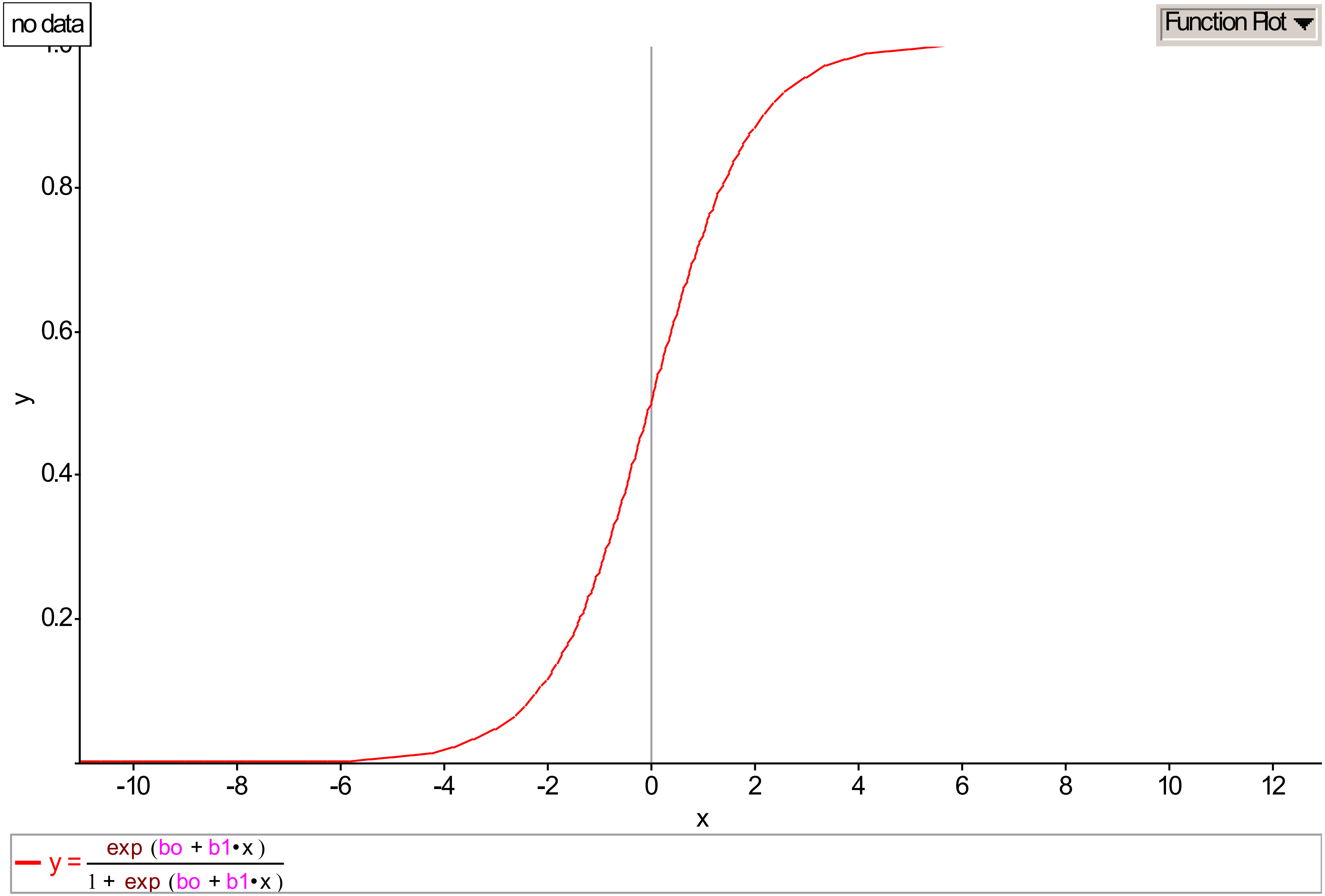
Here $\pi(x_i) = \frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)}, i = 1, \dots, n$

To make prediction (of Y), we threshold the estimated π with 0.5

If $\hat{\pi} \geq 0.5$, then $\hat{y} = 1$; If $\hat{\pi} < 0.5$, then $\hat{y} = 0$

Then you can generate the confusion matrix comparing y to $\hat{y}$

Logit Function



Binary Logistic Regression via R

```
> logitmodel=glm(Gender~Hgt,family=binomial,
data=Pulse)
> summary(logitmodel)

Call:
glm(formula = Gender ~ Hgt, family = binomial)

Deviance Residuals:
     Min        1Q     Median        3Q       Max
-2.77443  -0.34870  -0.05375   0.32973   2.37928

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14***
---
```

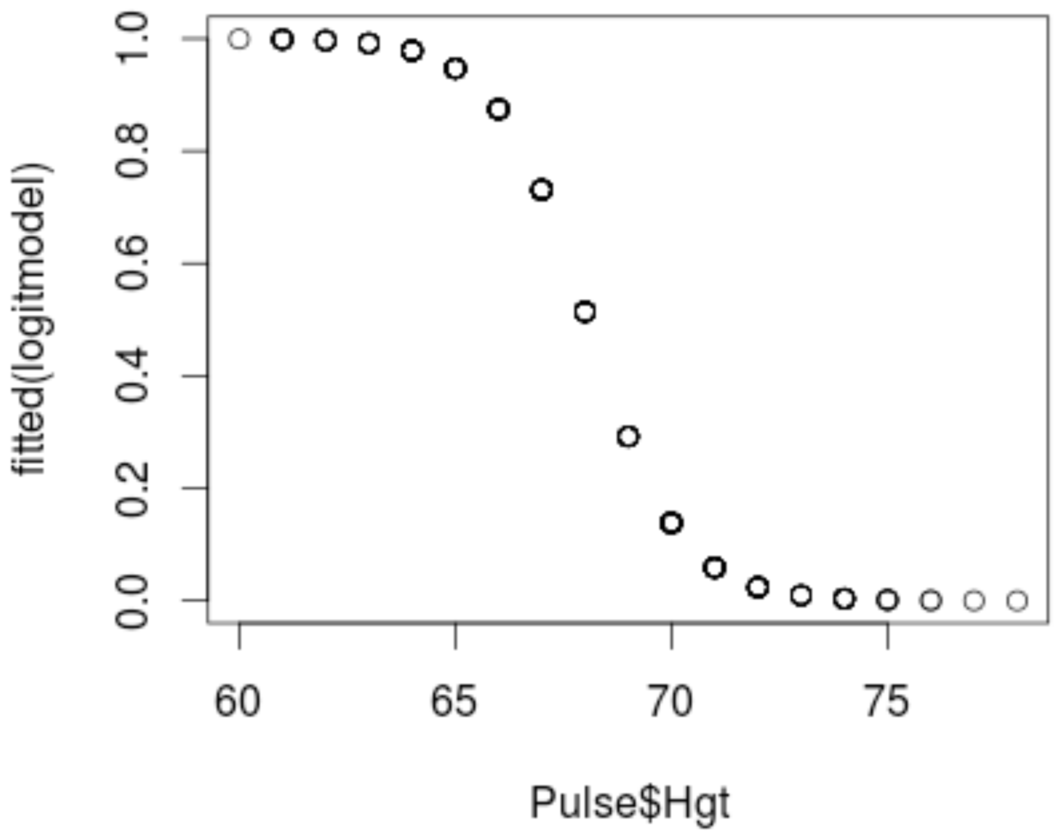
Note: this is the R command when we enter the data in the long form, that is, on a subject by subject basis.

```
Call:
glm(formula = Gender ~ Hgt, family = binomial, data = Pulse)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14***
---
```

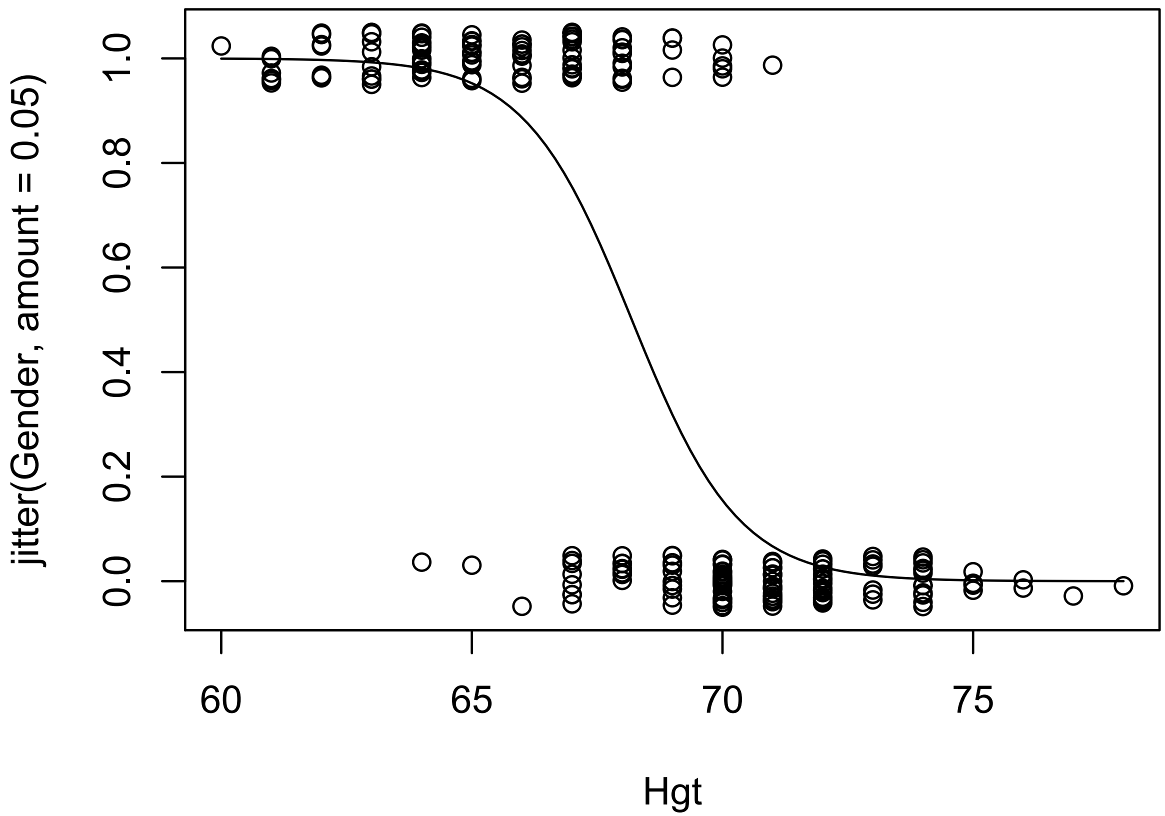
$$\hat{\pi} = \frac{e^{64.14 - 0.9424 * Hgt}}{1 + e^{64.14 - 0.9424 * Hgt}}$$

The estimated proportion of females (Gender = 1) in the population with height = Hgt;
It is also the estimated probabilty that a randomly selected subject from the population with height = Hgt is female.
Again, you see that the population share the same estimated model parameters, but you must use each subject's predictor value for his/her gender prediction, here being Hgt.

```
> plot(fitted(logitmodel)~Pulse$Hgt)
```



```
> with(Pulse,plot(Hgt,jitter(Gender,amount=0.05)))
> curve(exp(64.1-0.94*x)/(1+exp(64.1-0.94*x)), add=TRUE)
```



Dear students, by now, you have learned the basic concepts of the logistic regression and how to do it in R.

In the following slides:

(1) We shall provide additional examples, and, how to handle data in the short form (aka the summary data);

(2) We will also discuss the interpretation of the logistic model parameters in terms of the odds and odds ratio.

Example: Golf Putts

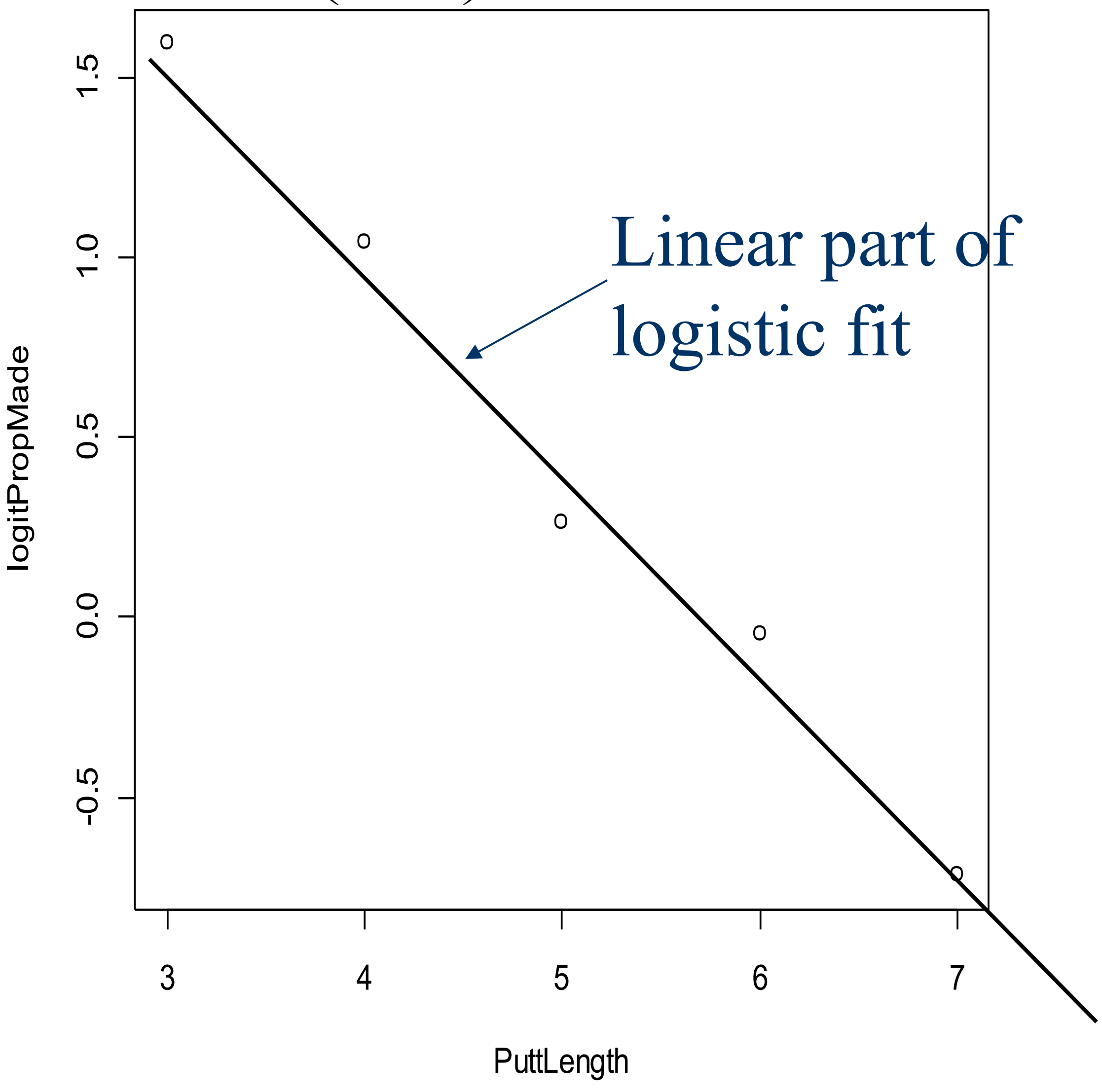
Length (x)	3	4	5	6	7
Made (y=1)	84	88	61	61	44
Missed (y=0)	17	31	47	64	90
Total	101	119	108	125	134

Build a model to predict the proportion of putts made (success) based on length (in feet).

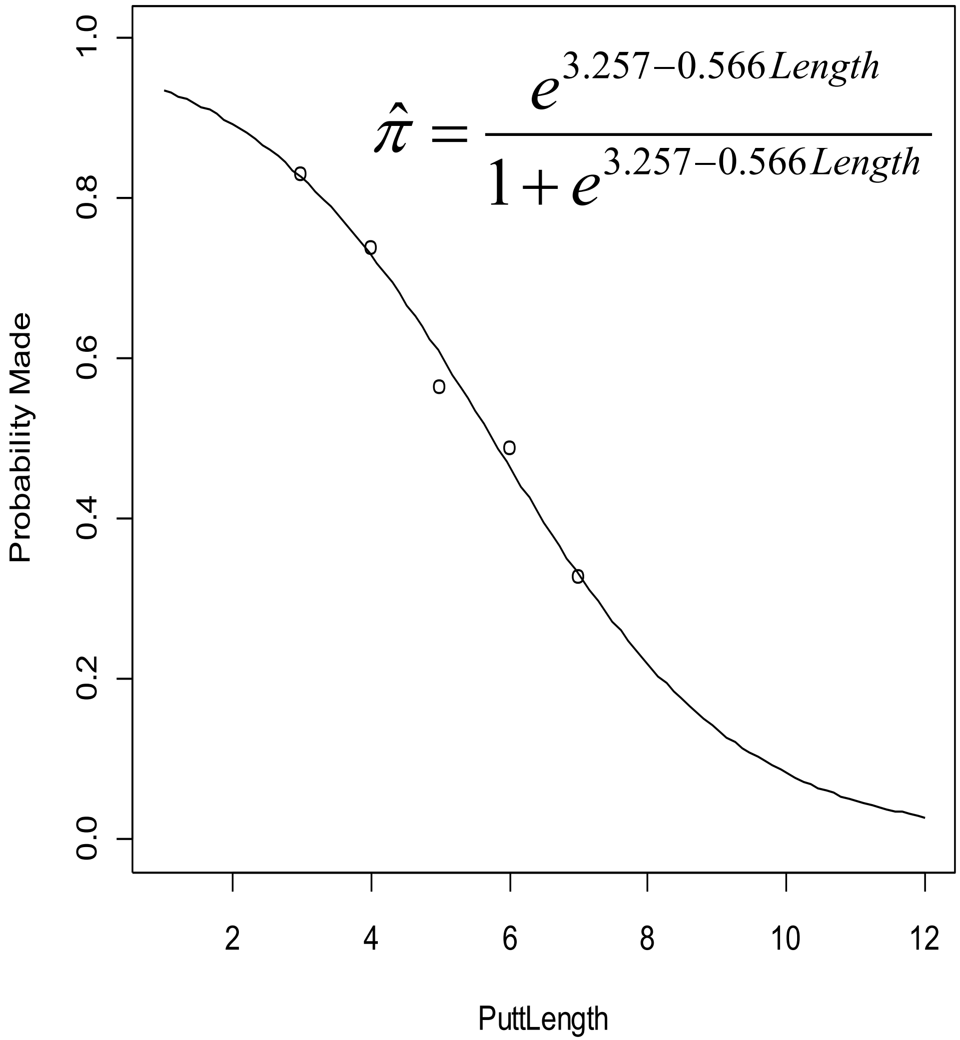
Logistic Regression for Putting

```
Call:
glm(formula = Made ~ Length, family = binomial, data =
Putts1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8705  -1.1186   0.6181   1.0026   1.4882

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.25684    0.36893   8.828   <2e-16 ***
Length      -0.56614    0.06747  -8.391   <2e-16 ***
---
```

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) \text{ vs. Length}$$



Probability Form of Putting Model



Odds

Definition:

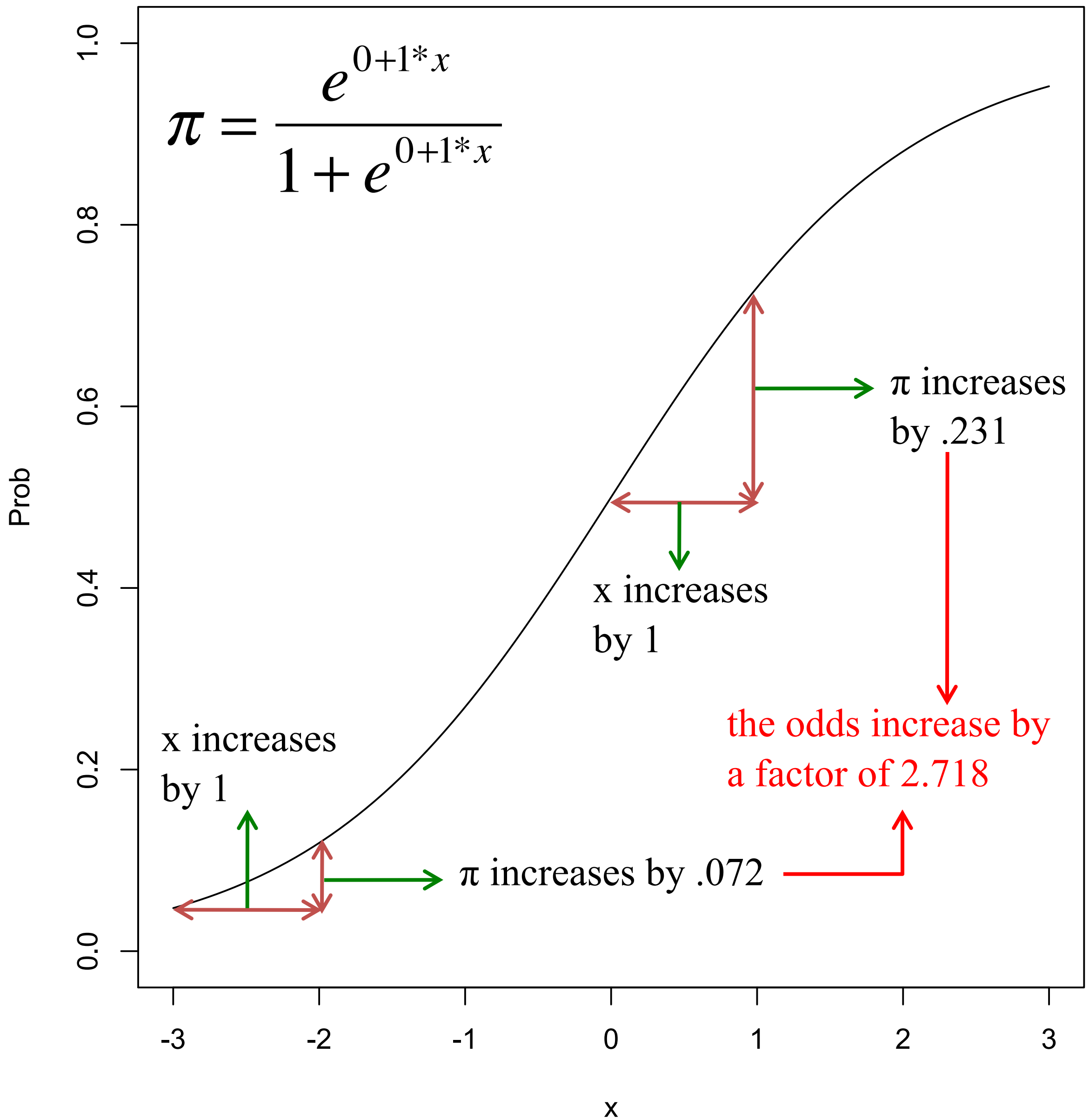
$$\frac{\pi}{1-\pi} = \frac{P(Yes)}{P(No)}$$ is the odds of Yes.

$$odds = \frac{\pi}{1-\pi} \Leftrightarrow \pi = \frac{odds}{1+odds}$$

$$\textbf{odds} = \frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}} \Leftrightarrow \boldsymbol{\pi} = \frac{\textbf{odds}}{1+\textbf{odds}}$$

Fair die

Event	Prob	Odds	
even #	1/2	1	[or 1:1]
X > 2	2/3	2	[or 2:1]
roll a 2	1/6	1/5	[or 1/5:1 or 1:5]



$\pi = \frac{e^{0+1*x}}{1+e^{0+1*x}}$
π increases
by .231
x increases
by 1
the odds increase by
a factor of 2.718
x increases
by 1
π increases by .072
Prob
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
x

Odds

Logit form of the model:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

$\Rightarrow$ The logistic model assumes a linear relationship between the *predictors* and *log(odds)*.

$$odds = \frac{\pi}{1-\pi} = e^{\beta_0 + \beta_1 X}$$

Odds Ratio

A common way to compare two groups is to look at the *ratio* of their odds

$$\text{Odds Ratio} = \text{OR} = \frac{\text{Odds}_1}{\text{Odds}_2}$$

Note: Odds ratio (OR) is similar to relative risk (RR).

$$\mathbf{RR} = \frac{\mathbf{p}_1}{\mathbf{p}_2} \qquad \mathbf{OR} = \mathbf{RR} * \frac{1-\mathbf{p}_2}{1-\mathbf{p}_1}$$

So when p is small, OR ≈ RR.

X is replaced by $X + 1$:

$$odds = e^{\beta_0+\beta_1 X}$$

is replaced by

$$odds = e^{\beta_0+\beta_1(X+1)}$$

So the ratio is

$$\frac{e^{\beta_0+\beta_1(X+1)}}{e^{\beta_0+\beta_1 X}} = e^{\beta_0+\beta_1(X+1)-(\beta_0+\beta_1 X)} = e^{\beta_1}$$

Example: TMS for Migraines

Transcranial Magnetic Stimulation vs. Placebo

Pain Free?	TMS	Placebo
YES	39	22
NO	61	78
Total	100	100

$$\hat{\pi}_{TMS} = 0.39 \qquad odds_{TMS} = \frac{39/100}{61/100} = \frac{39}{61} = 0.639 \qquad \hat{\pi} = \frac{0.639}{1+0.639} = 0.39$$

$$\hat{\pi}_{Placebo} = 0.22 \qquad odds_{Placebo} = \frac{22}{78} = 0.282$$

$$Odds\ ratio = \frac{0.639}{0.282} = 2.27$$

Odds are 2.27 times higher of getting relief using TMS than placebo

Logistic Regression for TMS data

```
> lmod=glm(cbind(Yes,No)~GroupTMS,family=binomial,data=TMS)
> summary(lmod)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2657     0.2414  -5.243 1.58e-07 ***
GroupTMS      0.8184     0.3167   2.584  0.00977 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.8854  on 1  degrees of freedom
Residual deviance: 0.0000  on 0  degrees of freedom
AIC: 13.701
```

Note: $e^{0.8184} = 2.27 =$ odds ratio

Note: Here we see how to use the glm function when
We have the short form the data.

```
Yes = c(39,22)    No = c(61,78)
GroupTMS = (1, 0)
```

Chi-Square Test for 2-way table

```
> datatable=rbind(c(39,22),c(61,78))
> datatable
     [,1] [,2]
[1,]   39   22
[2,]   61   78
> chisq.test(datatable,correct=FALSE)
    Pearson's Chi-squared test

data:  datatable
X-squared = 6.8168, df = 1, p-value = 0.00903
```

Binary Logistic Regression

```
> lmod=glm(cbind(Yes,No)~Group,family=binomial,data=TMS)
> summary(lmod)

Call:
glm(formula = cbind(Yes, No) ~ Group, family = binomial)Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2657     0.2414  -5.243 1.58e-07 ***
GroupTMS      0.8184     0.3167   2.584  0.00977 **
```

A Single *Binary Predictor* for a Binary Response

Response variable: Y = Success/Failure

Predictor variable: X = Group #1 / Group #2

- Method #1: Binary logistic regression
- Method #2: Z- test, compare two proportions
- Method #3: Chi-square test for 2-way table

All three "tests" are essentially equivalent, but the logistic regression approach allows us to mix other categorical and quantitative predictors in the model.

Putting Data

Odds using data from 6 feet = 0.953
Odds using data from 5 feet = 1.298

➔ Odds ratio (6 ft to 5 ft) = 0.953/1.298 = 0.73

The odds of making a putt from 6 feet are 73% of the odds of making from 5 feet.

Golf Putts Data

Length	3	4	5	6	7
Made	84	88	61	61	44
Missed	17	31	47	64	90
Total	101	119	108	125	134
$\hat{p}$	.8317	.7394	.5648	.4880	.3284
Odds	4.941	2.839	1.298	0.953	0.489

E.g., 5 feet: Odds $= \dfrac{.5648}{1-.5648} = \dfrac{61}{47} = 1.298$

E.g., 6 feet: Odds $= \dfrac{.4880}{1-.4880} = \dfrac{61}{64} = 0.953$

Golf Putts Data

Length	3	4	5	6	7
Made	84	88	61	61	44
Missed	17	31	47	64	90
Total	101	119	108	125	134
$\hat{p}$	.8317	.7394	.5648	.4880	.3284
Odds	4.941	2.839	1.298	.953	.489

OR	.575	.457	.734	.513

E.g., Odds $= \frac{.8317}{1-.8317} = \frac{84}{17} = 4.941$

Interpreting "Slope" using Odds Ratio

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

$$\Rightarrow \quad odds = e^{\beta_0 + \beta_1 X}$$

When we increase X by 1, the ratio of the new odds to the old odds is e^{β_1}.

i.e. odds are multiplied by e^{β_1}.

Odds Ratios for Putts

From samples at each distance:

4 to 3 feet	5 to 4 feet	6 to 5 feet	7 to 6 feet
0.575	0.457	0.734	0.513

From fitted logistic:

4 to 3 feet	5 to 4 feet	6 to 5 feet	7 to 6 feet
0.568	0.568	0.568	0.568

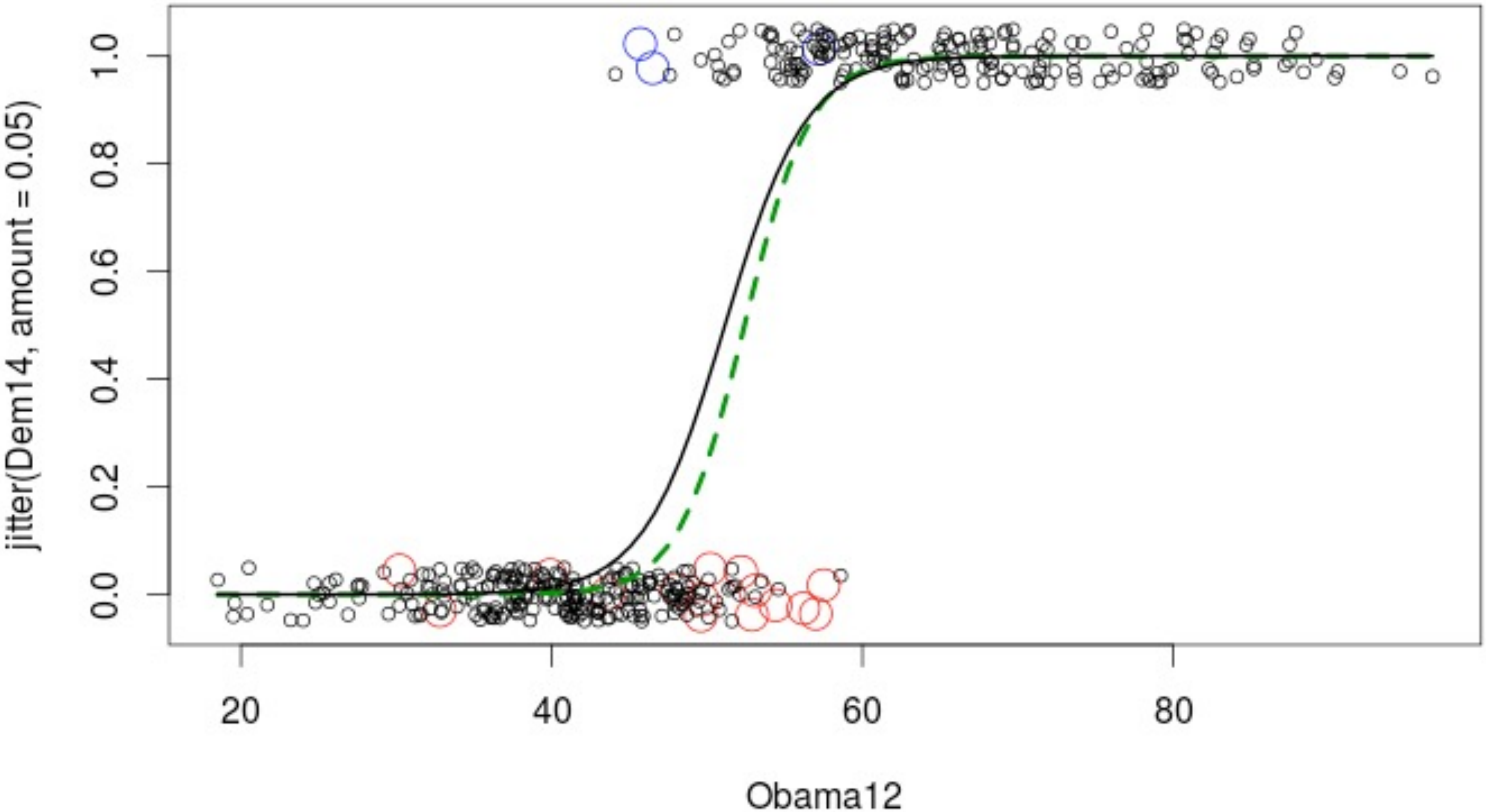
In a logistic model, the odds ratio is *constant* when changing the predictor by one.

Example: 2012 vs 2014 congressional elections

How does %vote won by Obama relate to a Democrat winning a House seat?

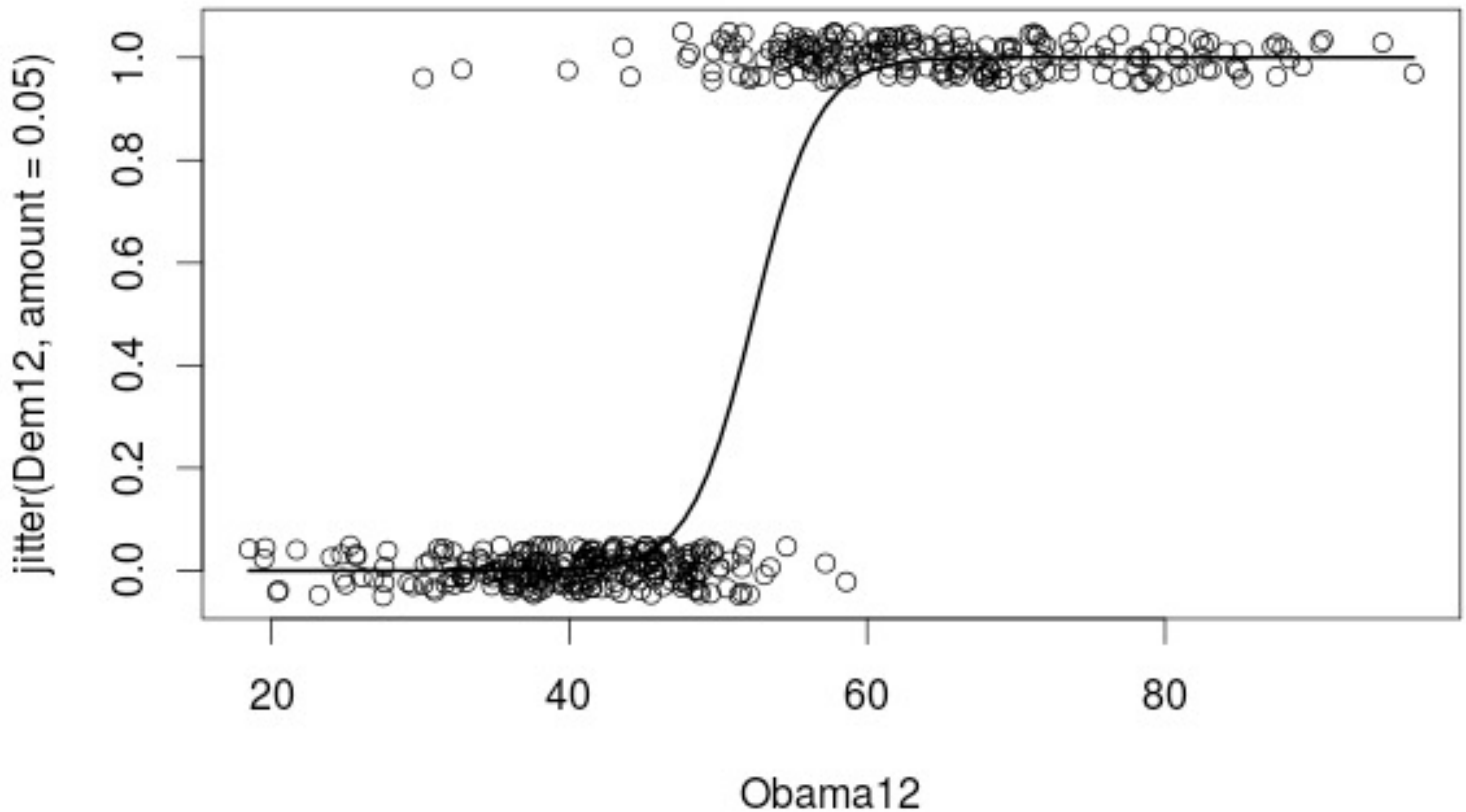
In 2012 a Democrat had a decent chance even if Obama got only 50% of the vote in the district. In 2014 that was less true.

In 2012 a Democrat had a decent chance even if Obama got only 50% of the vote in the district. In 2014 that was less true.



There is an easy way to graph logistic curves in R.

```
> library(TeachingDemos)
> with(elect, plot(Obama12,jitter(Dem12,amount=.05)))
> logitmod14=glm(Dem14~Obama12,family=binomial,data=elect)
> Predict.Plot(logitmod14, pred.var="Obama12",add=TRUE,
plot.args = list(lwd=3,col="black"))
```



R Logistic Output

```
> PuttModel=glm(Made~Length, family=binomial,data=Putts1)
> anova(PuttModel)
            Analysis of Deviance Table

          Df Deviance Resid. Df Resid. Dev
NULL                        586     800.21
Length     1   80.317       585     719.89
```

```
> summary(PuttModel)
Call:
glm(formula = Made ~ Length, family = binomial)
 Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.25684    0.36893   8.828   <2e-16 ***
Length      -0.56614    0.06747  -8.391   <2e-16 ***
---
    Null deviance: 800.21  on 586  degrees of freedom
Residual deviance: 719.89  on 585  degrees of freedom
```

Two forms of logistic data

1. Response variable Y = Success/Failure or 1/0: "long form" in which each case is a row in a spreadsheet (e.g., Putts1 has 587 cases). This is often called "binary response" or "Bernoulli" logistic regression.
2. Response variable Y = Number of Successes for a group of data with a common X value: "short form" (e.g., Putts2 has 5 cases – putts of 3 ft, 4 ft, … 7 ft). This is often called "Binomial counts" logistic regression.

Lengths	Makes	Misses	Trials
3	84	17	101
4	88	31	119
5	61	47	108
6	61	64	125
7	44	90	134

```
> str(Putts1)
'data.frame':  587 obs. of  2 variables:
 $ Length: int  3 3 3 3 3 3 3 3 3 3 ...
 $ Made  : int  1 1 1 1 1 1 1 1 1 1 ...
> longmodel=glm(Made~Length,family=binomial,data=Putts1)
> summary(longmodel)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.25684    0.36893   8.828   <2e-16 ***
Length      -0.56614    0.06747  -8.391   <2e-16 ***
---
    Null deviance: 800.21  on 586  degrees of freedom
Residual deviance: 719.89  on 585  degrees of freedom
```

Note: this is the R command when we enter the data in the long form, that is, on a subject by subject basis.

```
> str(Putts2)
'data.frame':  5 obs. of  4 variables:
 $ Length: int  3 4 5 6 7
 $ Made  : int  84 88 61 61 44
 $ Missed: int  17 31 47 64 90
 $ Trials: int  101 119 108 125 134
>
shortmodel=glm(cbind(Made,Missed)~Length,family=binomial,data=Putts2)
> summary(shortmodel)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.25684    0.36893   8.828   <2e-16 ***
Lengths     -0.56614    0.06747  -8.391   <2e-16 ***
---
    Null deviance: 81.3865  on 4  degrees of freedom
Residual deviance:  1.0692  on 3  degrees of freedom
```

Note: Here we see again how to use the glm function when we have the short form (summary) of the data.

Made = c(84, 88, 61, 61.44)

Missed = c(17, 31, 47, 64, 90)

Length = c(3, 4, 5, 6, 7)

Binary Logistic Regression Model

Y = Binary

X = Single predictor

π = proportion of 1's (yes, success) at any x

Equivalent forms of the logistic regression model:

Logit form $\log\left(\dfrac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$

Probability form $\pi = \dfrac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}}$

Binary Logistic Regression Model

Y = Binary | $X_1, X_2, \ldots, X_k$ = Multiple

π = proportion of 1's at any $x_1, x_2, \ldots, x_k$

Equivalent forms of the logistic regression model:

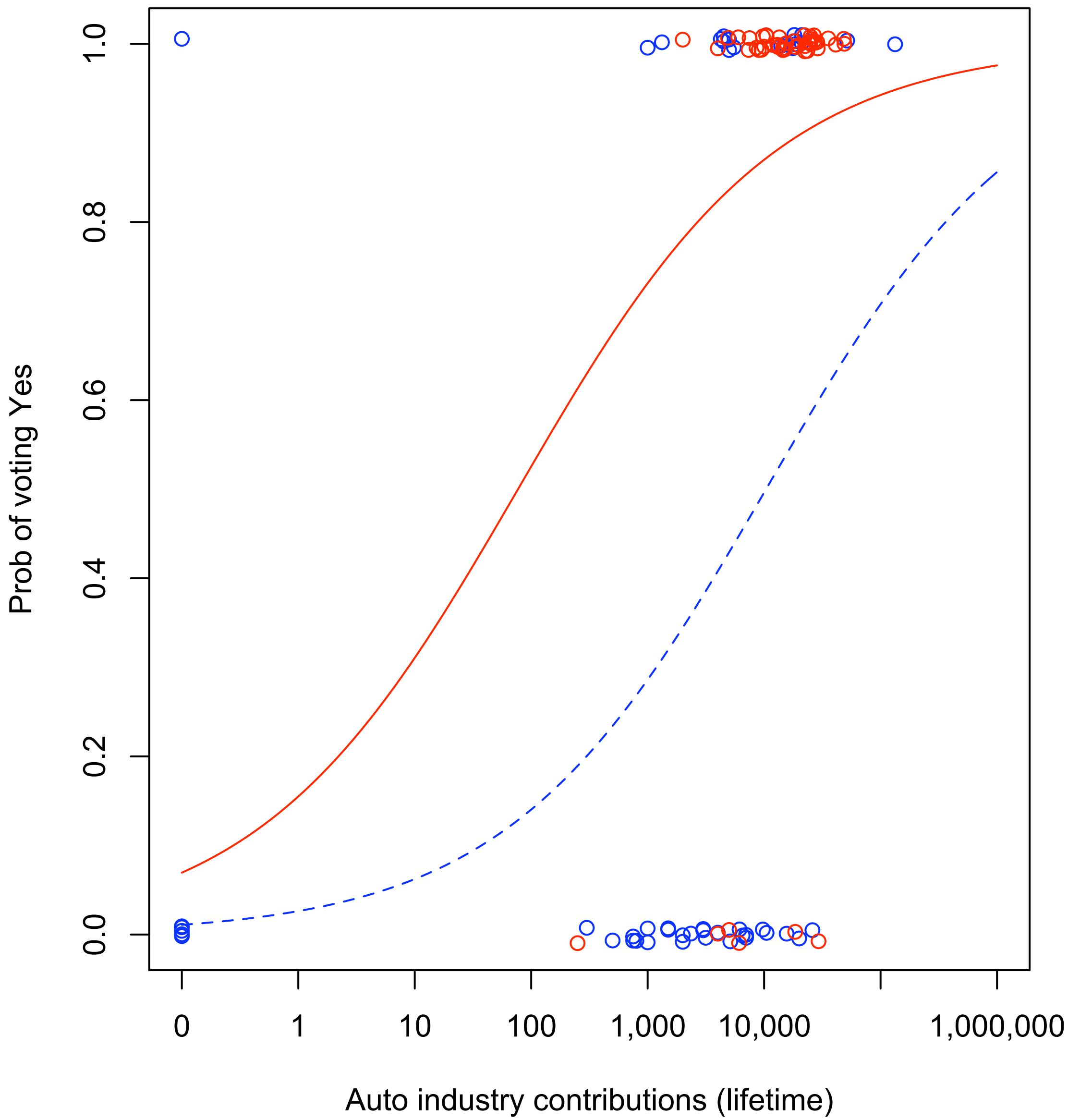
Logit form $$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Probability form $$\pi = \frac{e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}{1 + e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}$$

Interactions in logistic regression

Consider Survival in an ICU as a function of SysBP -- BP for short – and Sex

```
> intermodel=glm(Survive~BP*Sex, family=binomial, data=ICU)
> summary(intermodel)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.439304   1.021042  -1.410  0.15865
BP           0.022994   0.008325   2.762  0.00575 **
Sex          1.455166   1.525558   0.954  0.34016
BP:Sex      -0.013020   0.011965  -1.088  0.27653

    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 189.99  on 196  degrees of freedom
```



Rep = red,
Dem = blue

Lines are very close to parallel; not a significant interaction

Generalized Linear Model

(1) What is the link between Y and $\beta_0 + \beta_1 X$?

(a) General linear model: identity

(b) Logistic regression: logit

(c) Poisson regression: log

(2) What is the distribution of Y given X?

(a) General linear model : Normal (Gaussian)

(b) Logistic regression: Bernoulli

(c) Poisson regression: Poisson

C-index, a measure of concordance

Med school acceptance: predicted by MCAT and GPA?

Med school acceptance: predicted by coin toss??

```
> library(Stat2Data)
> data(MedGPA)
> str(MedGPA)
> GPA10=MedGPA$GPA*10
> Med.glm3=glm(Acceptance~MCAT+GPA10, family=binomial,
    data=MedGPA)
> summary(Med.glm3)
> Accept.hat <- Med.glm3$fitted > .5
> with(MedGPA, table(Acceptance,Accept.hat))
```

```
             Accept.hat
Acceptance  FALSE  TRUE
         0     18     7
         1      7    23
```

18 + 23 = 41 correct out of 55

```
> with(MedGPA, table(Acceptance,Accept.hat))
```

```
           Accept.hat
Acceptance  FALSE  TRUE
         0     18     7
         1      7    23
```

Now consider that there were 30 successes and 25 failures. There are 30*25=750 possible pairs.

We hope that the predicted Pr(success) is greater for the success than for the failure in a pair! If yes then the pair is “concordant”.

C-index = % concordant pairs

The R package rms has a command, lrm, that does logistic regression and gives the C-index.

```
> #C-index work using the MedGPA data
> library(rms) #after installing the rms package
> m3=lrm(Acceptance~MCAT+GPA10, data=MedGPA)
> m3
```

lrm(formula = Acceptance~ MCAT + GPA10)

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	55	LR chi2	21.78	R2	0.437	C	0.834
0	25	d.f.	2	g	2.081	Dxy	0.668
1	30	Pr(> chi2)	<0.0001	gr	8.015	gamma	0.669
max |deriv|	2e-07			gp	0.342	tau-a	0.337
				Brier	0. 167		

	Coef	S.E.	Wald Z	Pr(>|Z|)
Intercept	-22.373	6.454	-3.47	0.0005
MCAT	0.1645	0.1032	1.59	0.1108
GPA10	0.4678	0.1642	2.85	0.0044

Suppose we scramble the cases..

Then the C-index should be ½, like coin tossing

```
> newAccept=sample(MedGPA$Acceptance) #scramble the acceptances
> m1new=lrm(newAccept~MCAT+GPA10,data=MedGPA)
> m1new
```

lrm(formula = newAccept ~ MCAT + GPA10)

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	55	LR chi2	0.24	R2	0.006	C	0.520
0	25	d.f.	2	g	0.150	Dxy	0.040
1	30	Pr(> chi2)	0.8876	gr	1.162	gamma	0.041
max |deriv|	1e-13			gp	0.037	tau-a	0.020
				Brier	0.247		

	Coef	S.E.	Wald Z	Pr(>|Z|)
Intercept	-1.4763	3.4196	-0.43	0.6659
MCAT	0.0007	0.0677	0.01	0.9912
GPA10	0.0459	0.1137	0.40	0.6862

Important R Websites

1. **Logistic regression procedures, and how to split data into training and testing, and make predictions:**

http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/

2. **Stepwise and Best-subset variable selection methods using the information criteria (AIC or BIC):**

http://atm.amegroups.com/article/view/9706/pdf

I know they look tiny, but copy, paste and go, you will find them very helpful in your homework and exams, as always.

Acknowledgement

We thank Dr. Jeff Witmer for his wonderful examples!
We also thank colleagues who posted their notes on the internet.

Thank you!