Binary Logistic Regression Model

$Y =$ Binary response

$X =$ Quantitative predictor

$\pi =$ proportion of 1's (yes,success) at any X

Equivalent forms of the logistic regression model:

Logit form

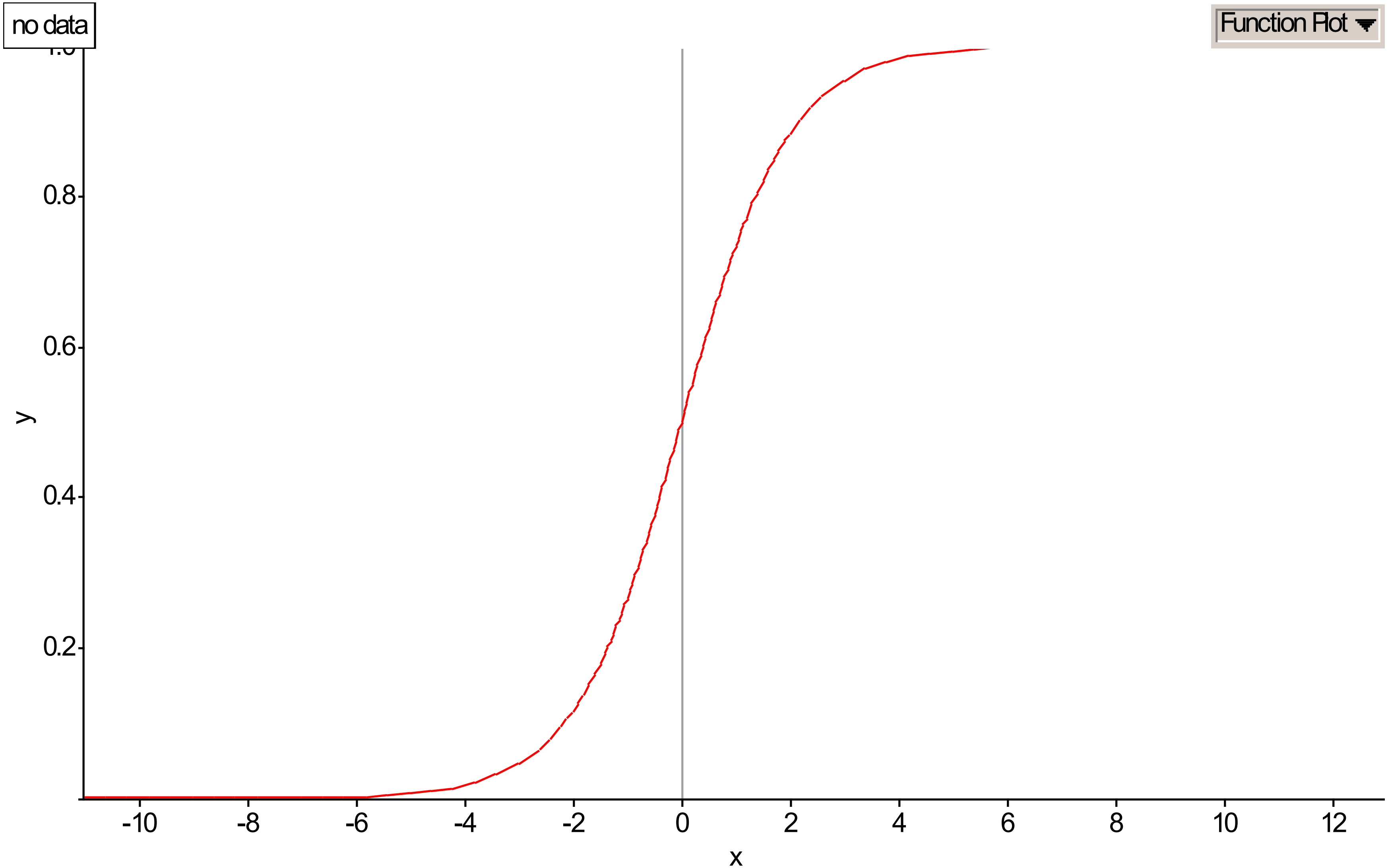
$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Probability form

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

N.B.: This is natural log (aka "ln")

Logit Function



$$y = \frac{\exp(bo + b1 \cdot x)}{1 + \exp(bo + b1 \cdot x)}$$

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***
---
```

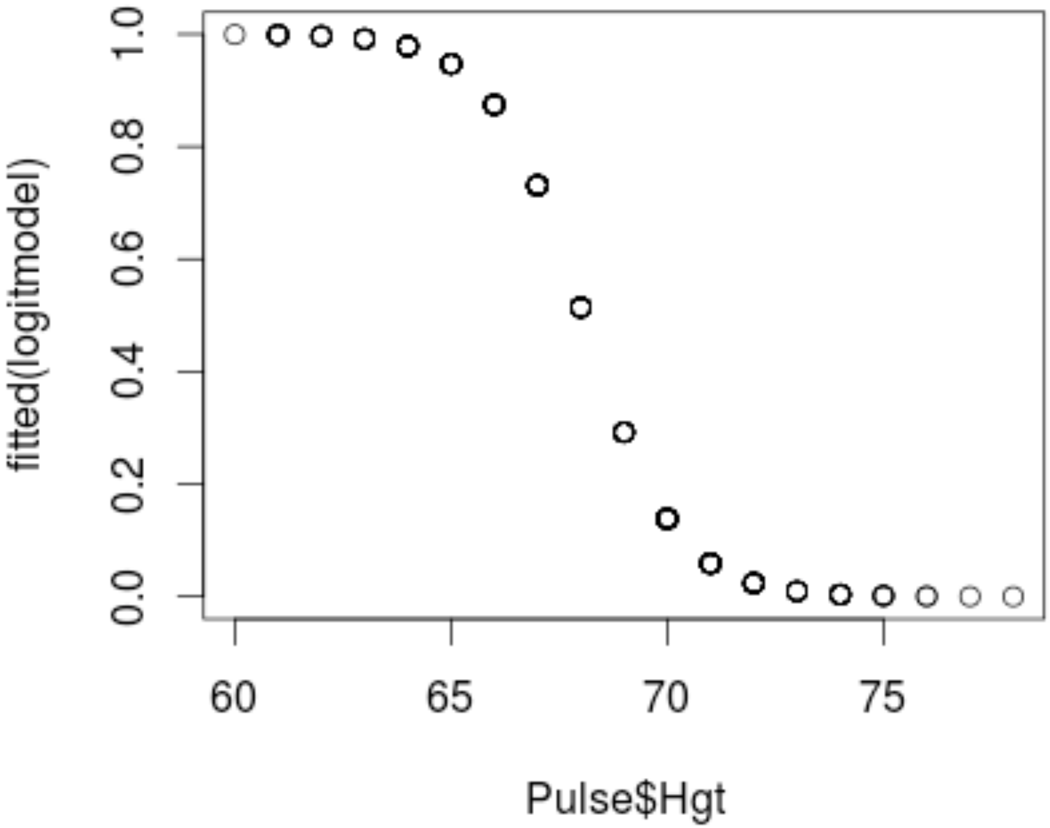
```
Call:
glm(formula = Gender ~ Hgt, family = binomial, data = Pulse)

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***
---
```

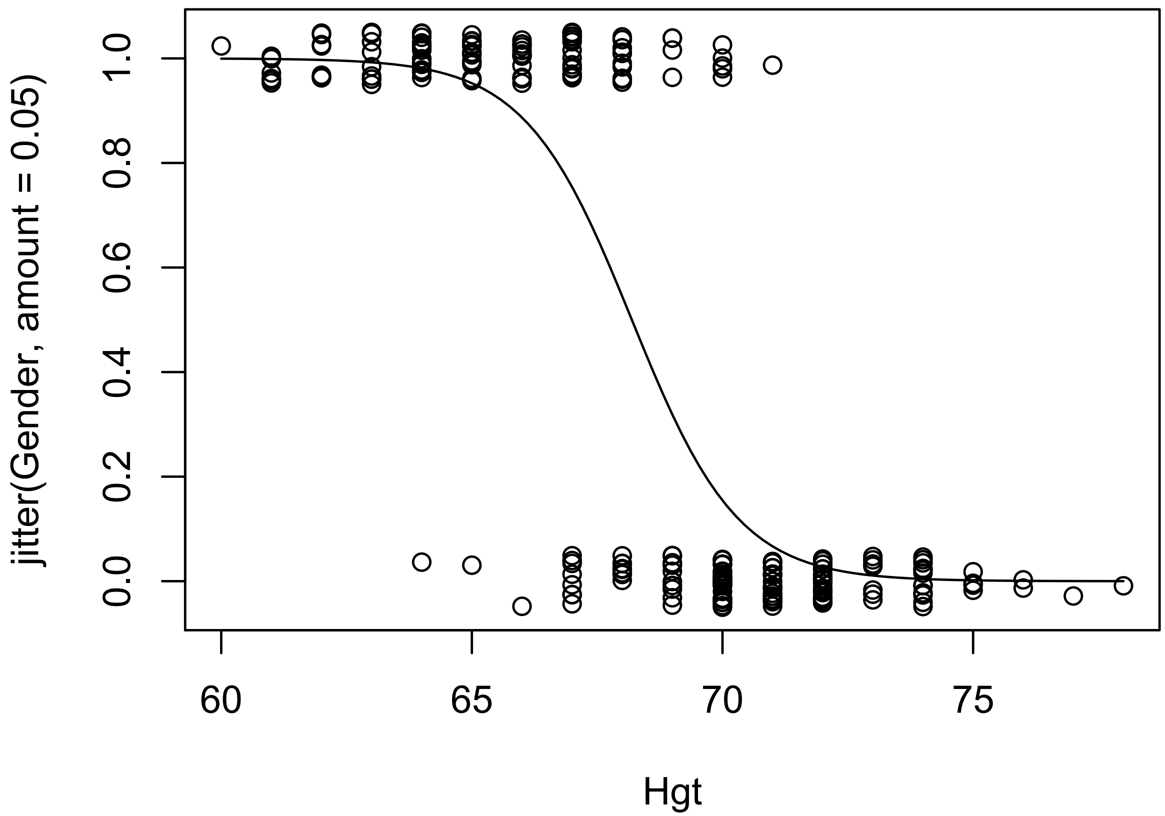
$$\hat{\pi} = \frac{e^{64.14 - 0.9424\,Ht}}{1 + e^{64.14 - .9424\,Ht}}$$

proportion of females at that 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)
```



Odds

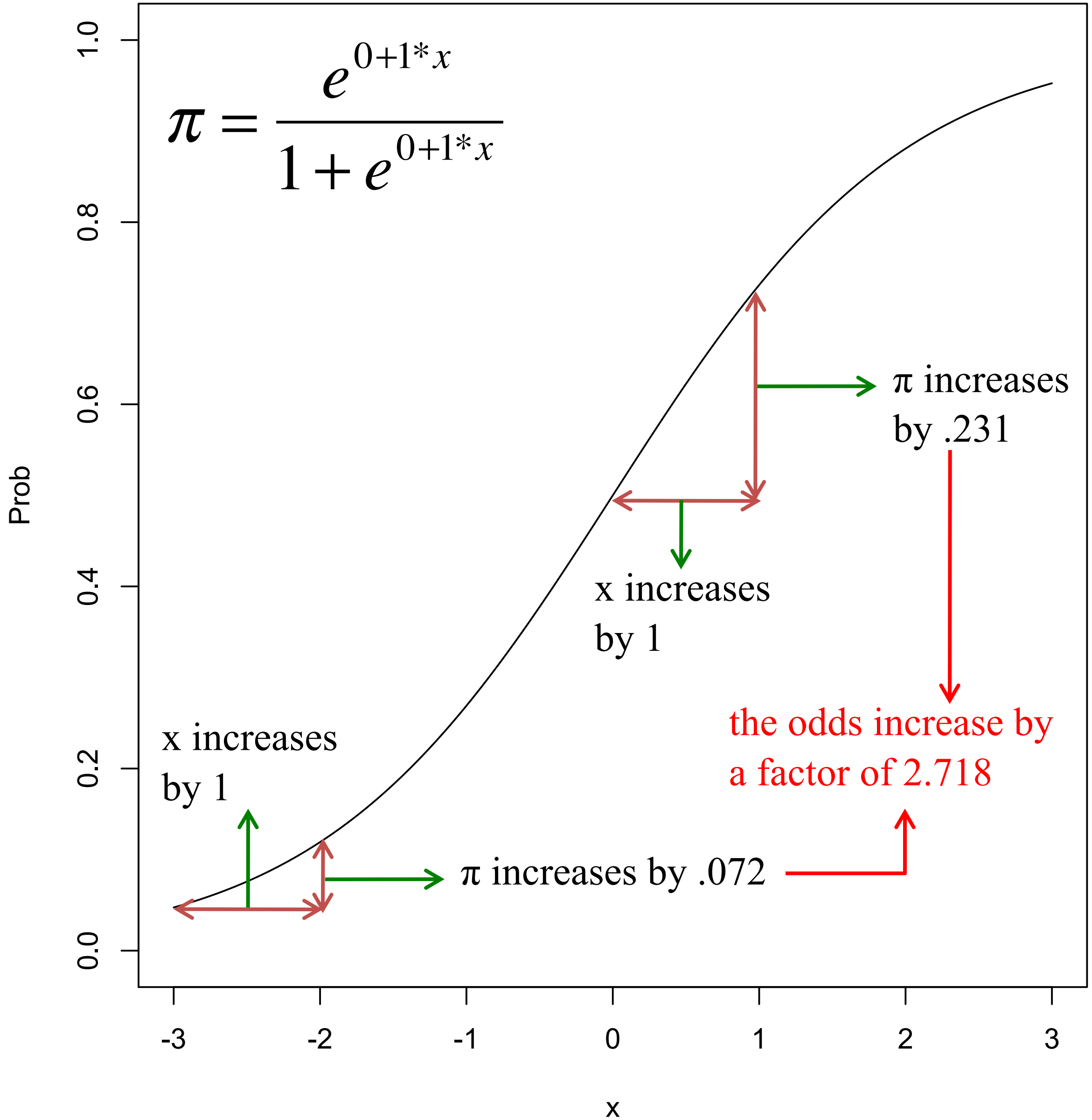
Definition:

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

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

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

Prob	Odds	
1/2	1	[or 1:1]
2/3	2	[or 2:1]
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
x
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3

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)*.

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

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

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

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

Example: Golf Putts

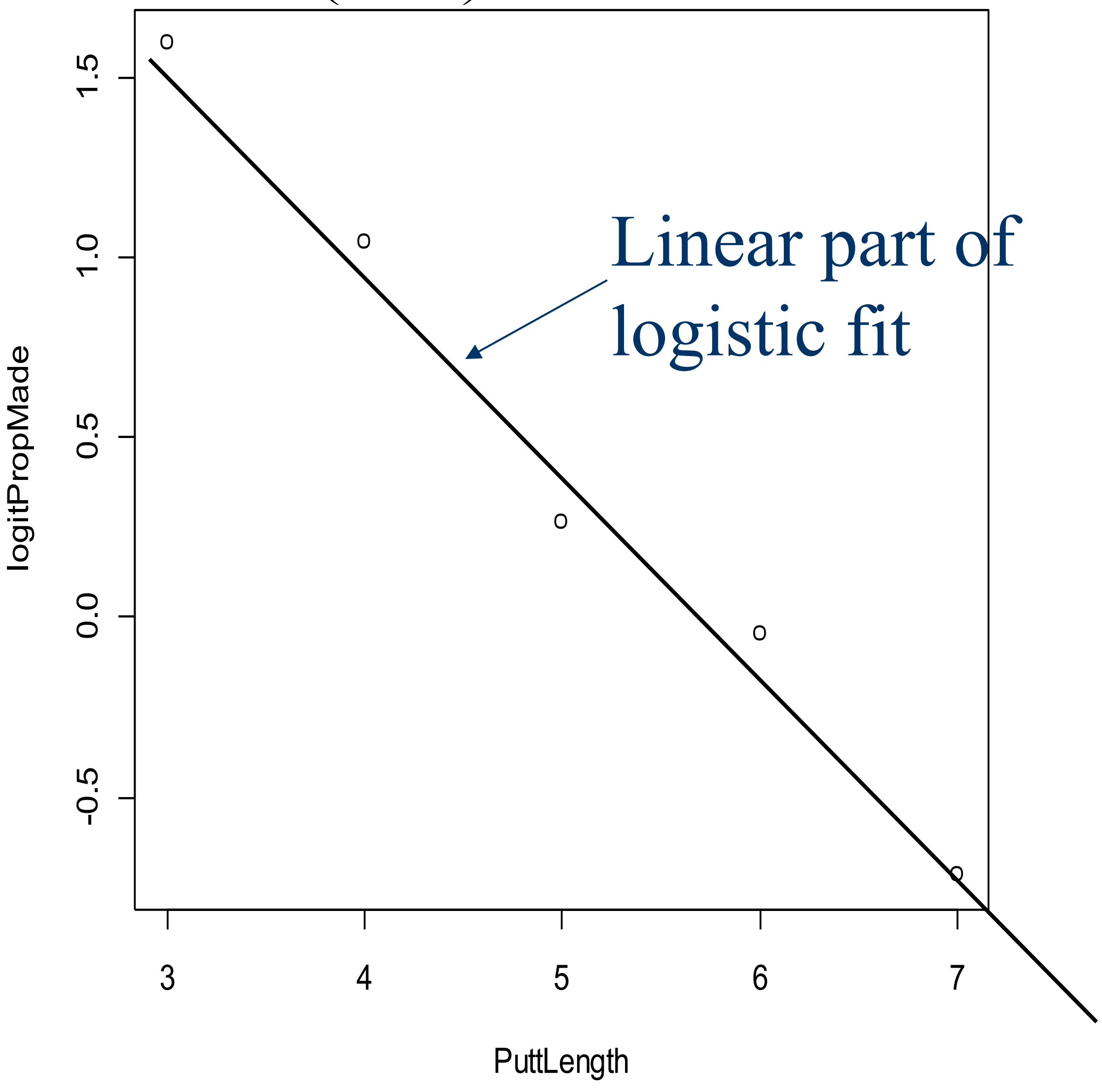
Length	3	4	5	6	7
Made	84	88	61	61	44
Missed	17	31	47	64	90
Total	101	119	108	125	134

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

```
> 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
```

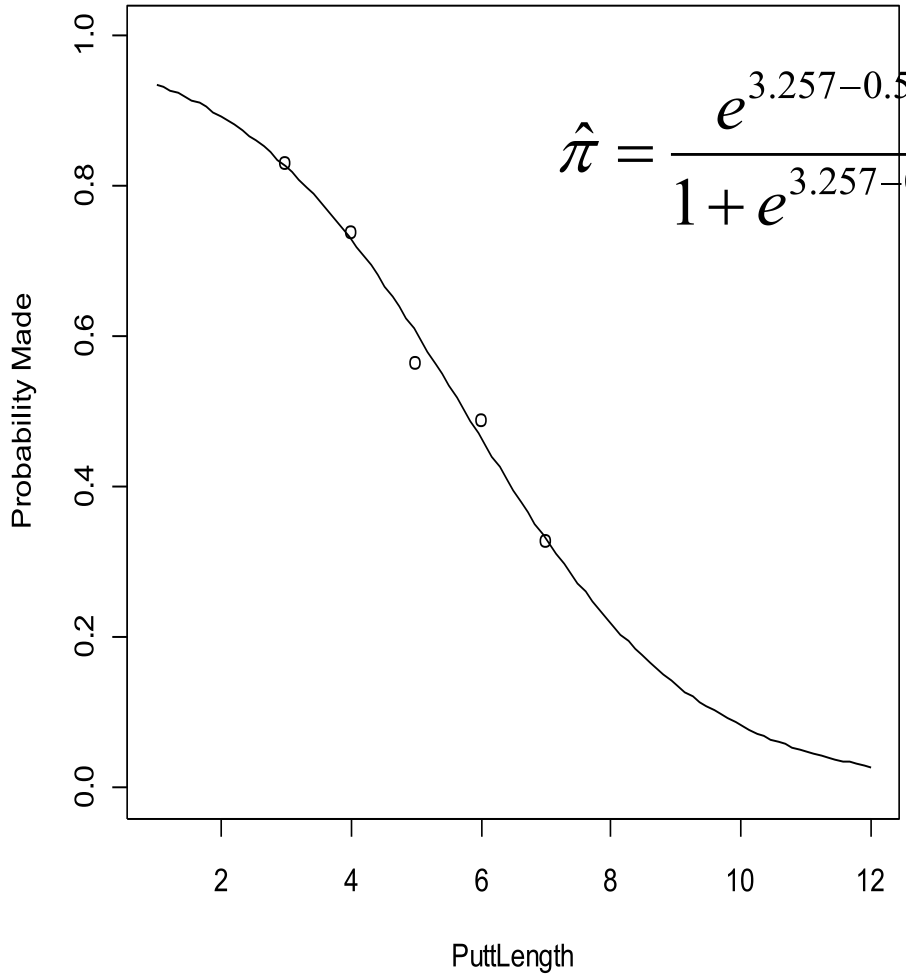
```
> str(Putts2)
'data.frame':  5 obs. of  4 variables:
 $ Length: int  3 4 5 6 7
 $ Made  : int  84 88 61 61 44
 $ Missed: int  17 31 47 64 90
 $ Trials: int  101 119 108 125 134
>
shortmodel=glm(cbind(Made,Missed)~Length,family=binomial,data=Putts2)
> summary(shortmodel)
Coefficients:
            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
```

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



Probability Form of Putting Model

```
>plot(PropMade~PuttLength,xlim=c(1,12),ylim=c(0,1))
>curve(exp(3.2568-.5661*x)/(1+exp(3.2568-.5661*x)),add=TRUE)
```



$$\hat{\pi} = \frac{e^{3.257-0.566\,Length}}{1+e^{3.257-0.566\,Length}}$$