Stat 306: Finding Relationships in Data.

Lecture 19

6.2 Logistic regression (Part 2)

From last lecture, we have three equivalent ways to write out the logistic regression model:

$$log\left(\frac{Pr(Y=1|\mathbf{X}=\mathbf{x})}{Pr(Y=0|\mathbf{X}=\mathbf{x})}\right)=\beta_0+\beta_1x_1+...+\beta_px_p$$

$$\frac{Pr(Y=1|\mathbf{X}=\mathbf{x})}{Pr(Y=0|\mathbf{X}=\mathbf{x})}=e^{\beta_0+\beta_1x_1+...+\beta_px_p}$$

$$Pr(Y=1|\mathbf{X}=\mathbf{x})=\frac{e^{\beta_0+\beta_1x_1+...+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+...+\beta_px_p}}$$

Recall that: k is the number of β's (here $p+1$)

From last lecture, we have three equivalent ways to write out the logistic regression model:

log-odds

$$log\left(\frac{Pr(Y=1|\mathbf{X}=\mathbf{x})}{Pr(Y=0|\mathbf{X}=\mathbf{x})}\right)=\beta_0+\beta_1x_1+...+\beta_px_p$$

odds

$$\frac{Pr(Y=1|\mathbf{X}=\mathbf{x})}{Pr(Y=0|\mathbf{X}=\mathbf{x})}=e^{\beta_0+\beta_1x_1+...+\beta_px_p}$$

probability

$$Pr(Y=1|\mathbf{X}=\mathbf{x})=\frac{e^{\beta_0+\beta_1x_1+...+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+...+\beta_px_p}}$$

Recall that: k is the number of β's (here $p+1$)

Odds:

$$\frac{Pr(Y=1|\mathbf{X}=\mathbf{x})}{Pr(Y=0|\mathbf{X}=\mathbf{x})} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$$

Let's call

$$\pi_1 = Pr(Y=1|\mathbf{X}=\mathbf{x}_1)$$

$$\pi_2 = Pr(Y=1|\mathbf{X}=\mathbf{x}_2)$$

Then :

$$odds_1 = \frac{\pi_1}{(1-\pi_1)}$$ and $$odds_2 = \frac{\pi_2}{(1-\pi_2)}$$

Odds:

$$\frac{Pr(Y=1|\mathbf{X}=\mathbf{x})}{Pr(Y=0|\mathbf{X}=\mathbf{x})} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$$

Let’s call

$$\pi_1 = Pr(Y=1|\mathbf{X}\boxed{=\mathbf{x}_1)}$$

Examples:

Placebo

vs.

Treatment

Non-Smoking

vs.

Smoking

$$\pi_2 = Pr(Y=1|\mathbf{X}\boxed{=\mathbf{x}_2)}$$

Then :

$$odds_1 = \frac{\pi_1}{(1-\pi_1)}$$ and $$odds_2 = \frac{\pi_2}{(1-\pi_2)}$$

probability

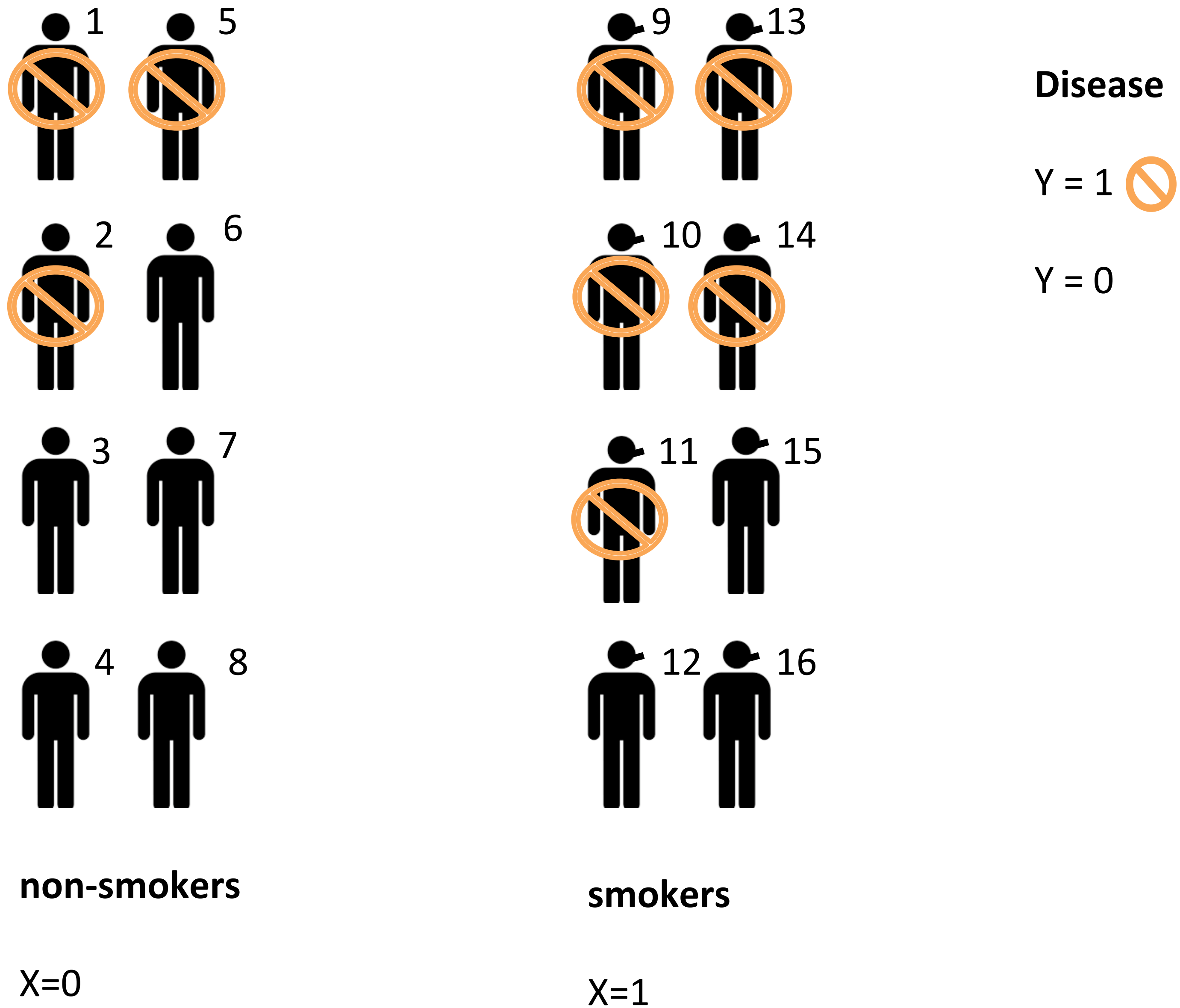
$$\pi_1 = Pr(Y=1|\mathbf{X}=\mathbf{x}_1) \quad \text{and} \quad \pi_2 = Pr(Y=1|\mathbf{X}=\mathbf{x}_2)$$

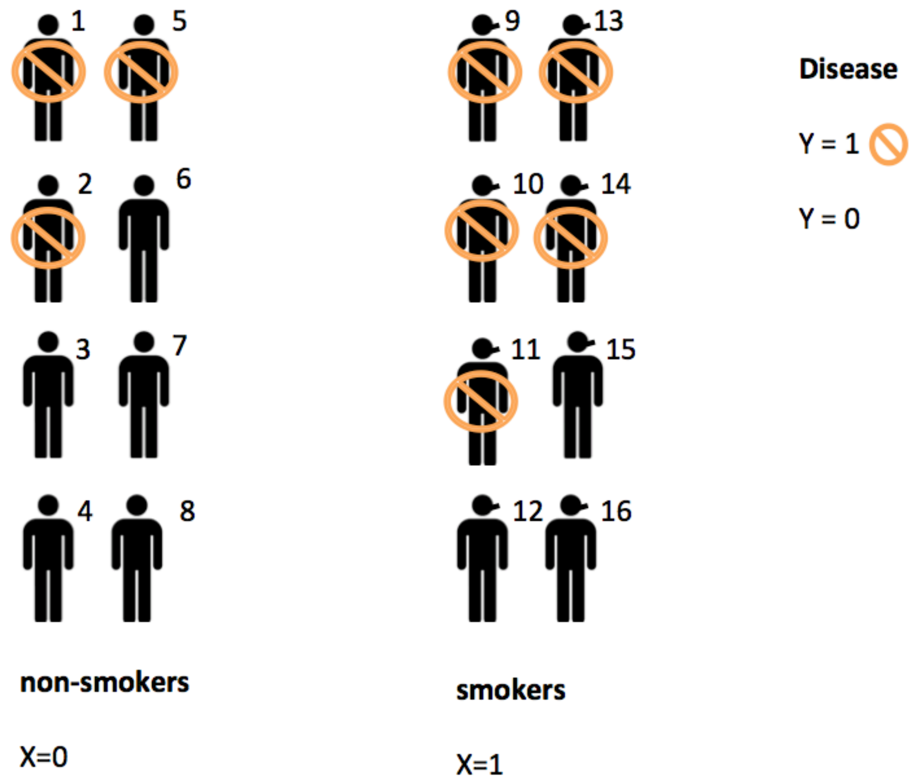
odds

$$odds_1 = \frac{\pi_1}{(1-\pi_1)} \quad \text{and} \quad odds_2 = \frac{\pi_2}{(1-\pi_2)}$$

Odds Ratio

$$OR = \frac{\frac{\pi_1}{(1-\pi_1)}}{\frac{\pi_2}{(1-\pi_2)}}$$


1
5
2
6
3
7
4
8
9
13
10
14
11
15
12
16
Disease
Y = 1
Y = 0
non-smokers
X=0
smokers
X=1

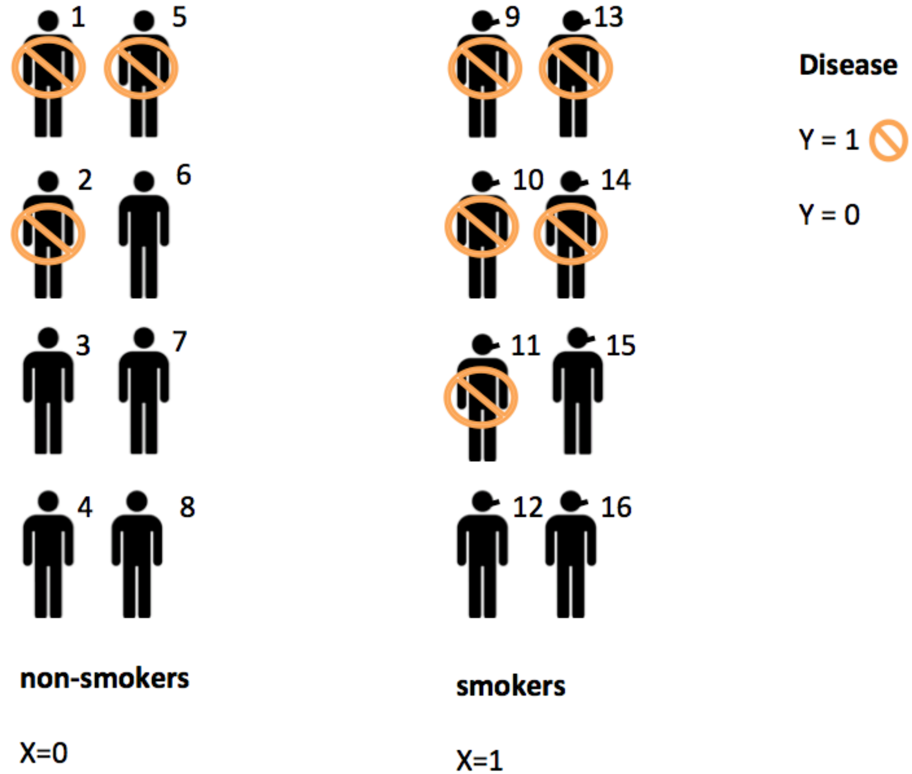


probability $\Pr(Y=1|X=0) = 3/8 = 0.375$ $\Pr(Y=1|X=0) = 5/8 = 0.625$

odds $\text{odds}_{X=0} = 3/5 = 0.6$ $\text{odds}_{X=1} = 5/3 = 1.667$

ODDS RATIO $(5/3) / (3/5) = 25/9 = 2.778$

Interpretation: The odds of being diseased are 2.778 times higher for smokers than for non-smokers.

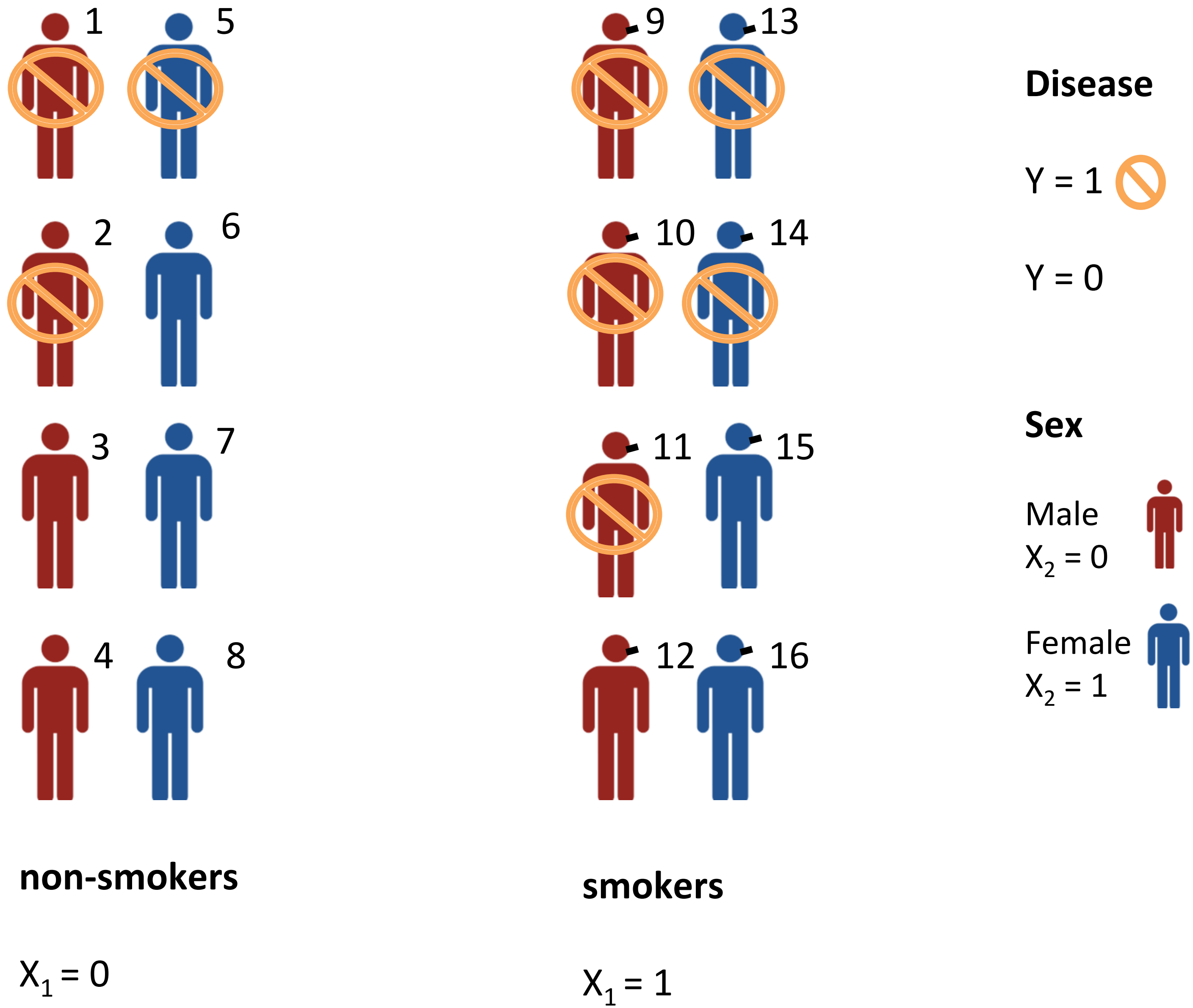


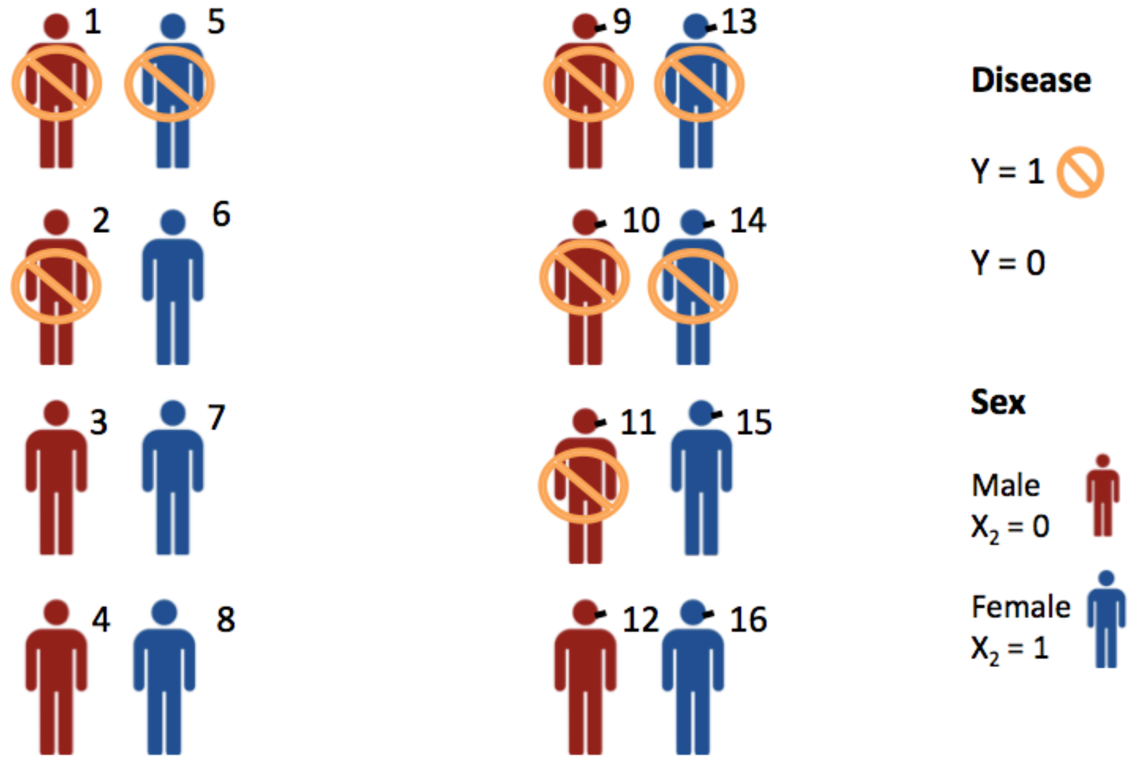
probability	Pr(Y=1 | X=0) = 3/8 = 0.375	Pr(Y=1 | X=0) = 5/8 = 0.625
odds	$\text{odds}_{X=0}$ = 3/5 = 0.6	$\text{odds}_{X=1}$ = 5/3 = 1.667

Exercise: Recall the logistic model: $log\left(\frac{Pr(Y=1|\mathbf{X}=\mathbf{x})}{Pr(Y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1$ .... Therefore: OR = exp(β_1)

ODDS RATIO (5/3) / (3/5) = 25/9 = 2.778

Interpretation: The odds of being diseased are 2.778 times higher for smokers than for non-smokers.


1
5
9
13
Disease
Y = 1
2
6
10
14
Y = 0
3
7
11
15
Sex
Male
X2 = 0
4
8
12
16
Female
X2 = 1
non-smokers
X1 = 0
smokers
X1 = 1



non-smokers $X_1 = 0$

smokers $X_1 = 1$

probability

$Pr(Y=1|X_1=0, X_2=0) = 2/4 = 0.5$

$Pr(Y=1|X_1=0, X_2=1) = 1/4 = 0.25$

$Pr(Y=1|X_1=1, X_2=0) = 3/4 = 0.75$

$Pr(Y=1|X_1=1, X_2=1) = 2/4 = 0.5$

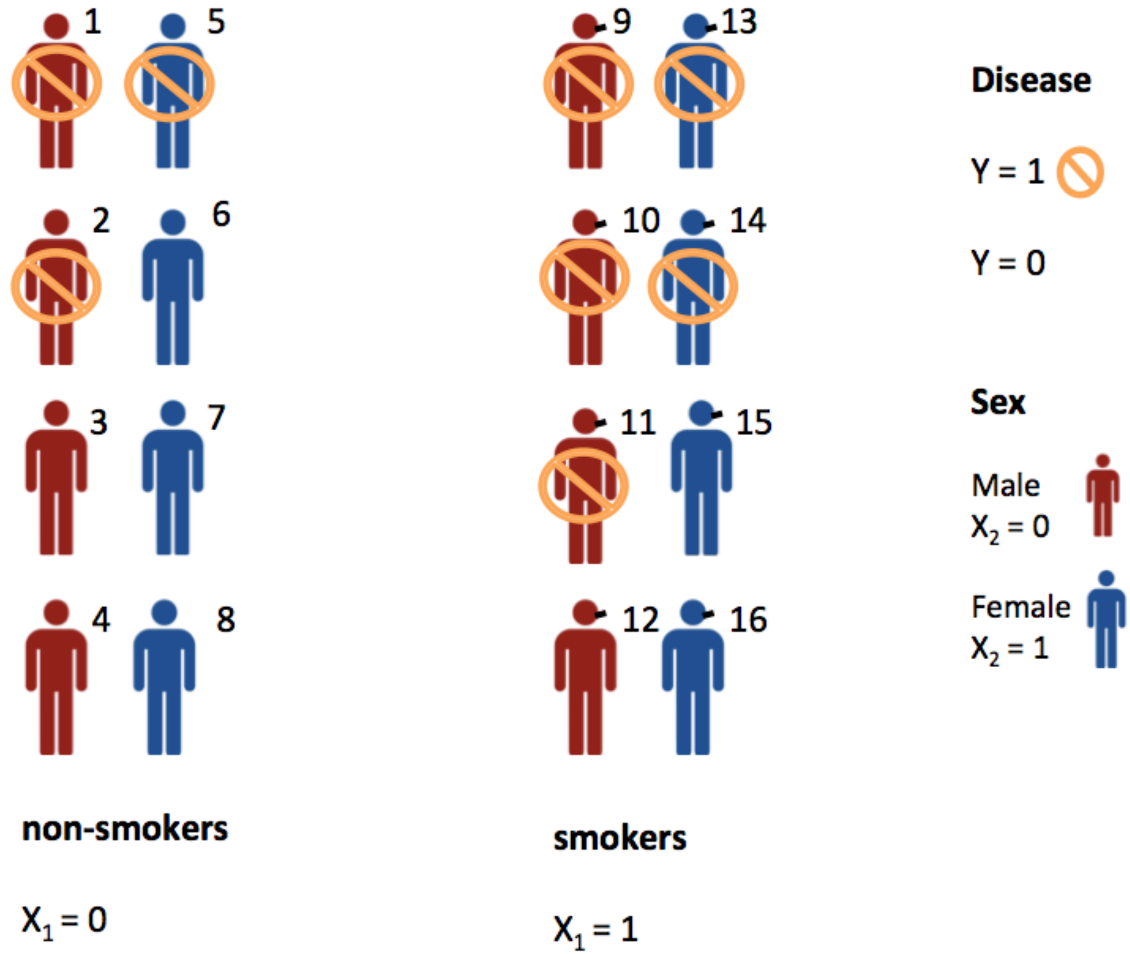
odds

$odds_{X1=0,\ X2=0} = 2/2 = 1{:}1$

$odds_{X1=0,\ X2=1} = 1/3 = 0.333$

$odds_{X1=1,\ X2=0} = 3/1 = 3$

$odds_{X1=1,\ X2=1} = 2/2 = 1$



probability

$\Pr(Y=1|X_1=0, X_2=0) = 2/4 = 0.5$

$\Pr(Y=1|X_1=0, X_2=1) = 1/4 = 0.25$

$\Pr(Y=1|X_1=1, X_2=0) = 3/4 = 0.75$

$\Pr(Y=1|X_1=1, X_2=1) = 2/4 = 0.5$

odds

$\text{odds}_{X1=0,\ X2=0} = 2/2 = 1{:}1$

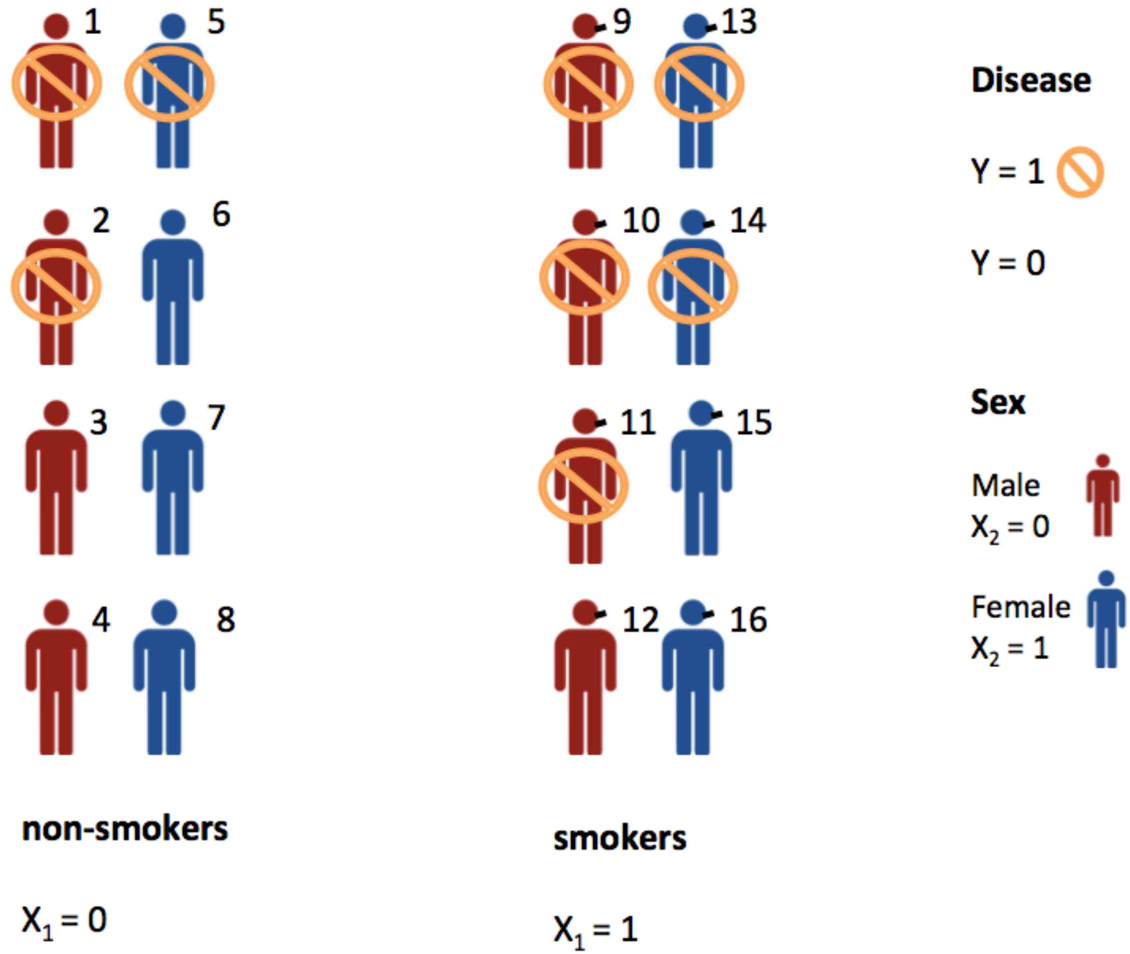
$\text{odds}_{X1=0,\ X2=1} = 1/3 = 0.333$

$\text{odds}_{X1=1,\ X2=0} = 3/1 = 3$

$\text{odds}_{X1=1,\ X2=1} = 2/2 = 1$

odds ratio

OR_{male}: 3:1 OR_{female}: 3:1



probability

$\Pr(Y=1 | X_1=0, X_2=0) = 2/4 = 0.5$
$\Pr(Y=1 | X_1=0, X_2=1) = 1/4 = 0.25$

$\Pr(Y=1 | X_1=1, X_2=0) = 3/4 = 0.75$
$\Pr(Y=1 | X_1=1, X_2=1) = 2/4 = 0.5$

odds

$\text{odds}_{X1=0,\ X2=0} = 2/2 = 1{:}1$

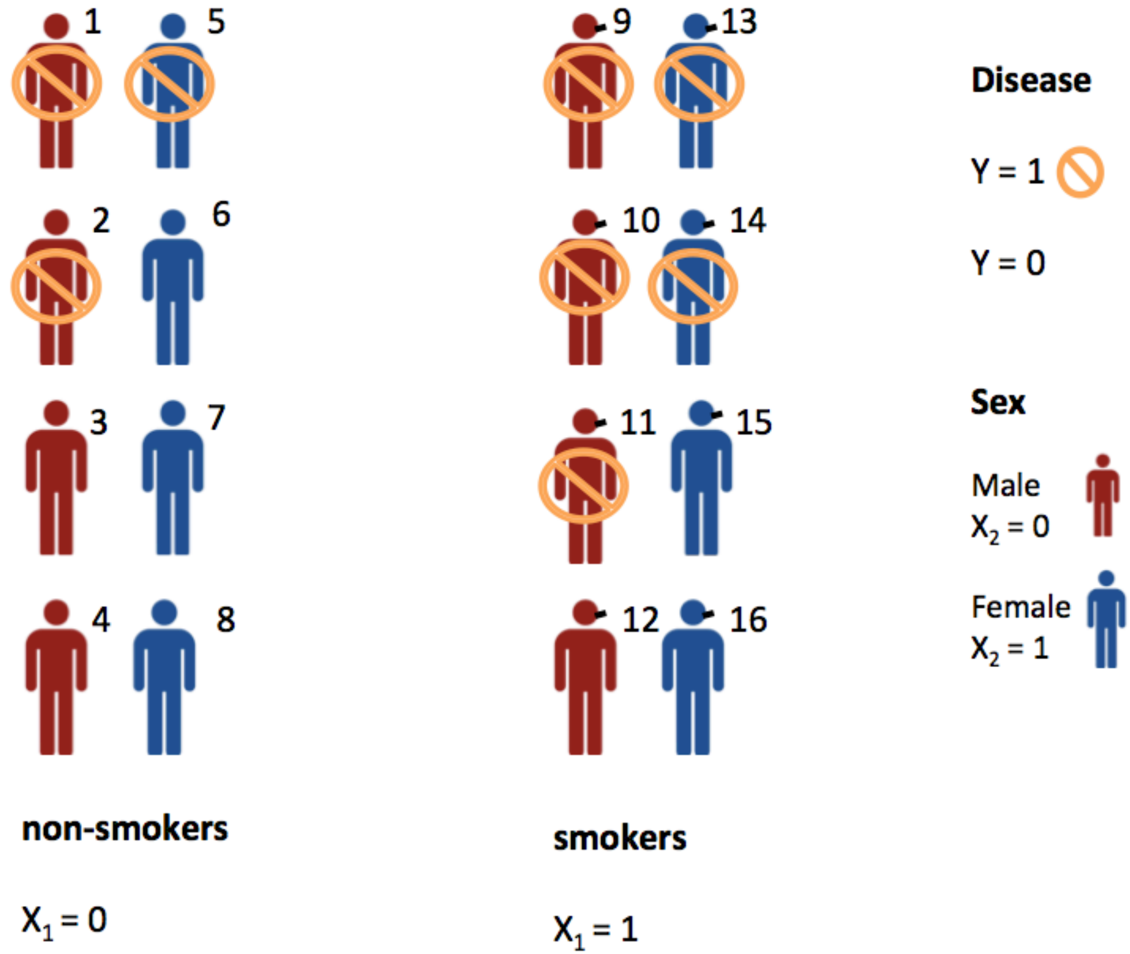
$\text{odds}_{X1=0,\ X2=1} = 1/3 = 0.333$

$\text{odds}_{X1=1,\ X2=0} = 3/1 = 3$

$\text{odds}_{X1=1,\ X2=1} = 2/2 = 1$

odds ratio

OR_{male}: 3:1 OR_{female}: 3:1 .... so do we have OR = 3 or OR = 2.778 ?



odds ratio $\quad OR_{male}$: 3:1 $\quad OR_{female}$: 3:1 $\quad$ .... so do we have OR = 3 or OR = 2.778 ?

Notes:

- This has nothing to do with unobserved confounding, same thing happens if we randomize subjects.
  (Exercise : Check to see that $cor(x_1, x_2) = 0$)
- This happens because of the “non-collapsibility” of the OR.

Reading : "Second argument: Omitted non-confounders in logistic regression"
http://jakewestfall.org/blog/index.php/2018/03/12/logistic-regression-is-not-fucked/

In both these example the covariate "X" is uncorrelated with the covariate "Colour"

Yet, since the logistic function is not collapsible, the average of the blue and red curves is not equal to the black curve…. very curious…

- Maybe there are better measure to describe the effect?
- Since OR is so difficult to interpret, perhaps we should use RR?

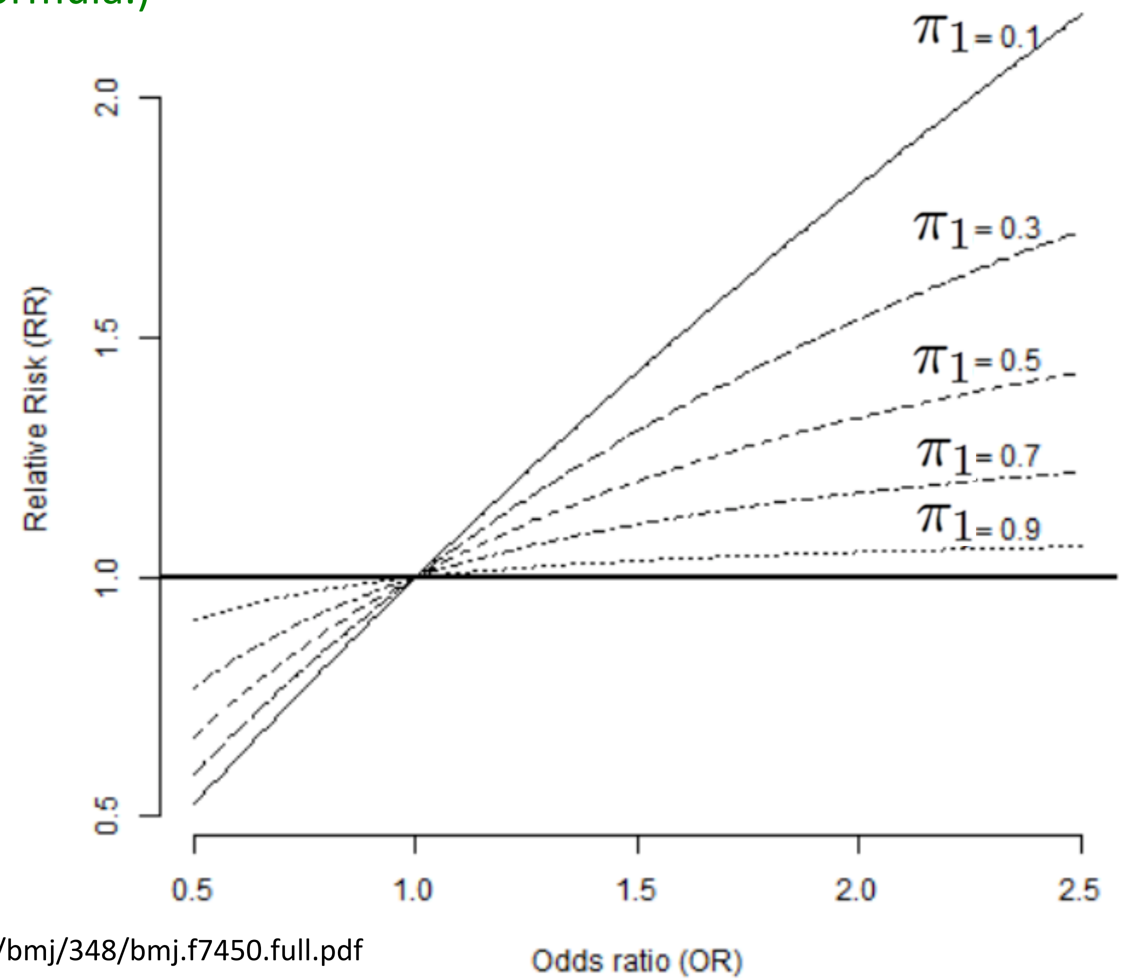
Type θ	*Expression*	*Domain*	*Null Value*
Risk difference (RD)	$\pi_1 - \pi_2$	$[-1, 1]$	0
Relative risk (RR)	π_1/π_2	$(0, \infty)$	1
log RR	$\log(\pi_1) - \log(\pi_2)$	$(-\infty, \infty)$	0
Odds ratio (OR)	$\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$	$(0, \infty)$	1
log OR	$\log \frac{\pi_1}{1-\pi_1} - \log \frac{\pi_2}{1-\pi_2}$	$(-\infty, \infty)$	0

Biostatistical Methods: The Assessment of Relative Risks
By John M. Lachin

To convert an Odds Ratio to a Relative Risk, you need to know π_1, which in our example is Pr(Y=1|X=0). Here is the formula:

$$RR = OR/(1 - \pi_1 + (\pi_1 \cdot OR))$$

(Exercise : Derive the formula.)



More info: http://www.bmj.com/content/bmj/348/bmj.f7450.full.pdf

Maximum Likelihood with 5 coins tosses…

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.59	0.33	0.17	0.08	0.03	0.01	0.00	0.00	0.00
1	0.33	0.41	0.36	0.26	0.16	0.08	0.03	0.01	0.00
2	0.07	0.20	0.31	0.35	0.31	0.23	0.13	0.05	0.01
3	0.01	0.05	0.13	0.23	0.31	0.35	0.31	0.20	0.07
4	0.00	0.01	0.03	0.08	0.16	0.26	0.36	0.41	0.33
5	0.00	0.00	0.00	0.01	0.03	0.08	0.17	0.33	0.59

$$Prob(\pi|N,k) = \binom{N}{k} \cdot \pi^k (1-k)^{N-k}$$

Maximum Likelihood with 5 coins tosses…

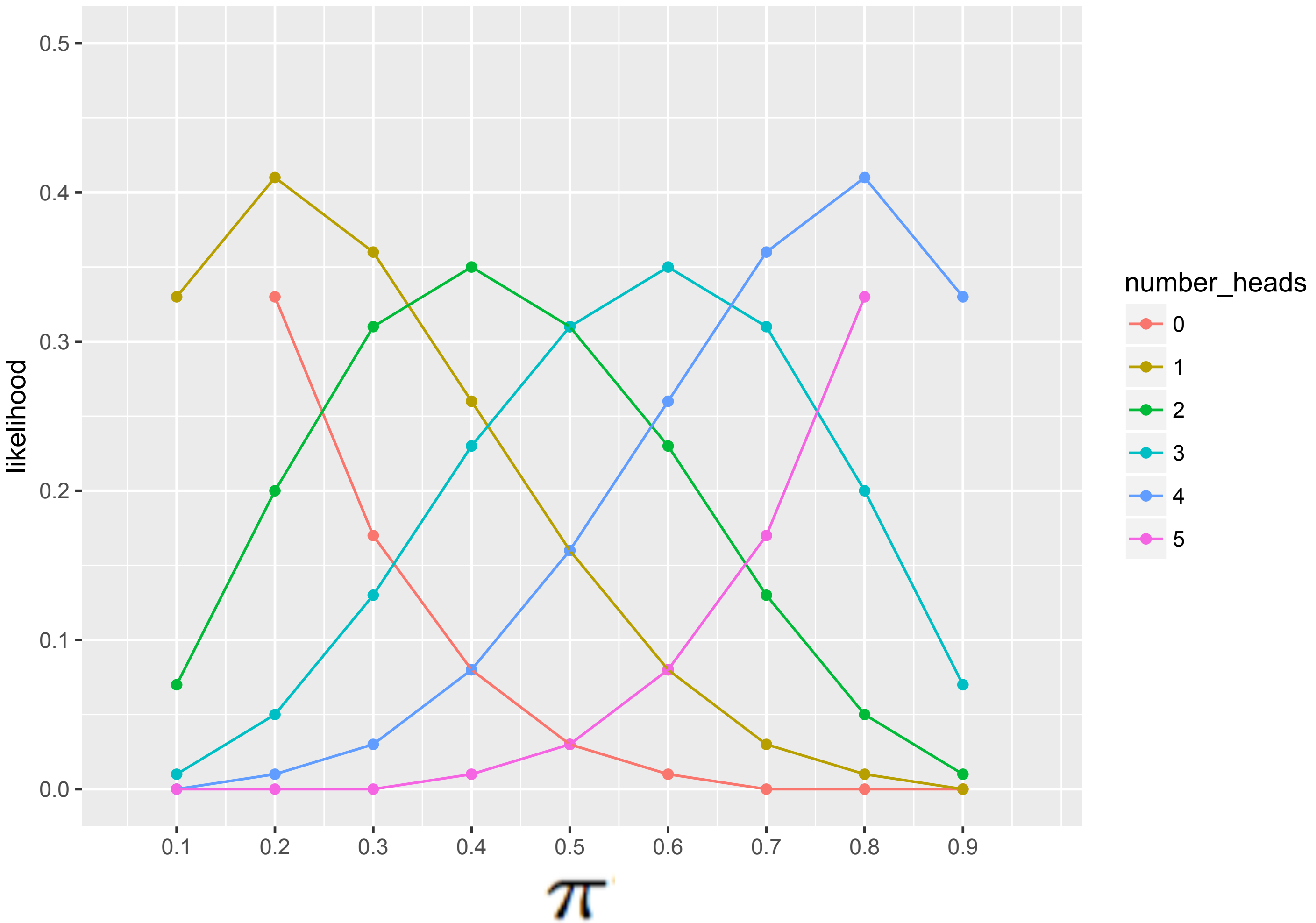
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.59	0.33	0.17	0.08	0.03	0.01	0.00	0.00	0.00
1	0.33	0.41	0.36	0.26	0.16	0.08	0.03	0.01	0.00
2	0.07	0.20	0.31	0.35	0.31	0.23	0.13	0.05	0.01
3	0.01	0.05	0.13	0.23	0.31	0.35	0.31	0.20	0.07
4	0.00	0.01	0.03	0.08	0.16	0.26	0.36	0.41	0.33
5	0.00	0.00	0.00	0.01	0.03	0.08	0.17	0.33	0.59

Examples:

Prob(π = 0.5 | 0 heads out of 5 tosses) = 0.03

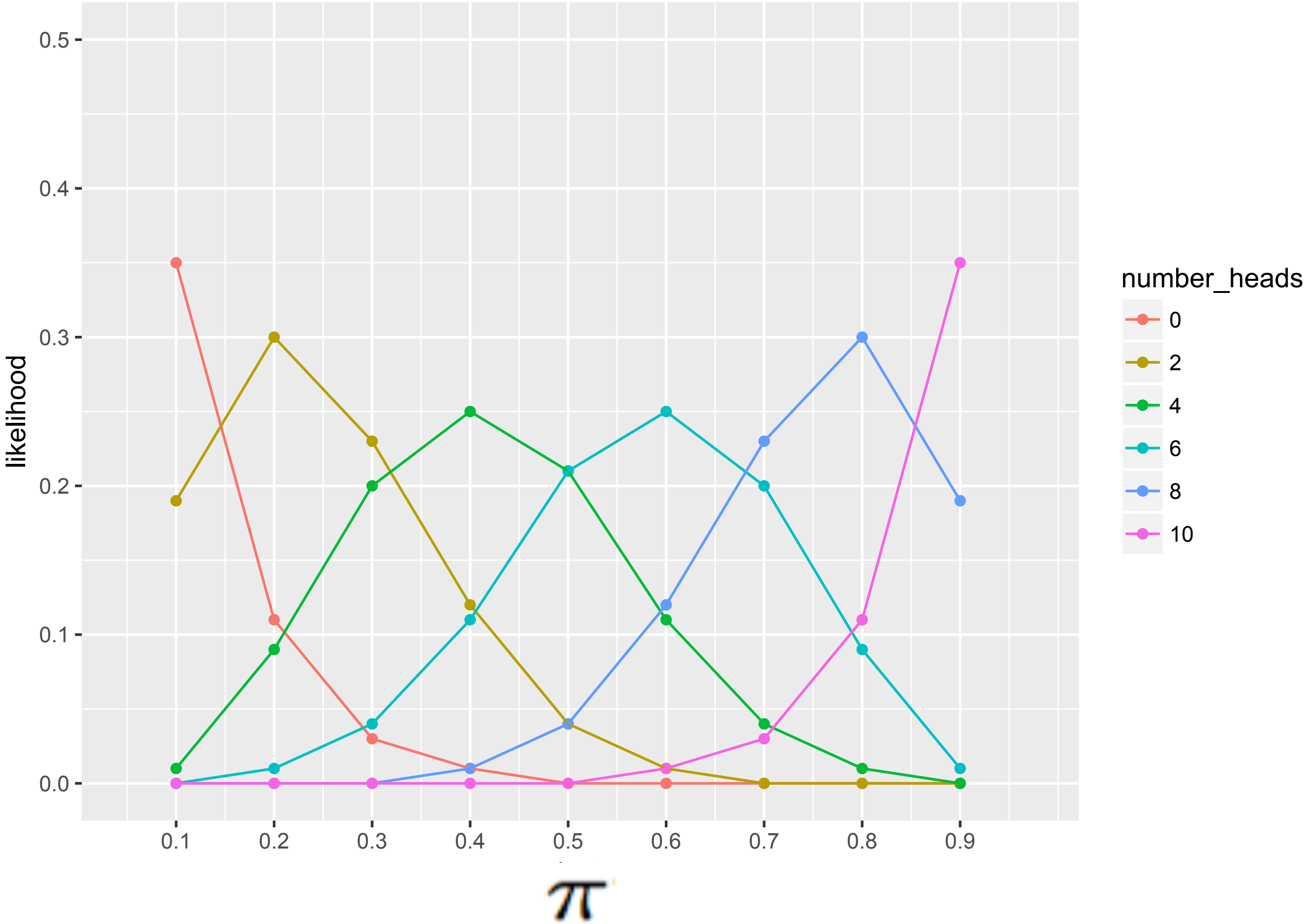
Prob(π = 0.2 | 4 heads out of 5 tosses) = 0.01

Maximum Likelihood with 5 coins tosses…



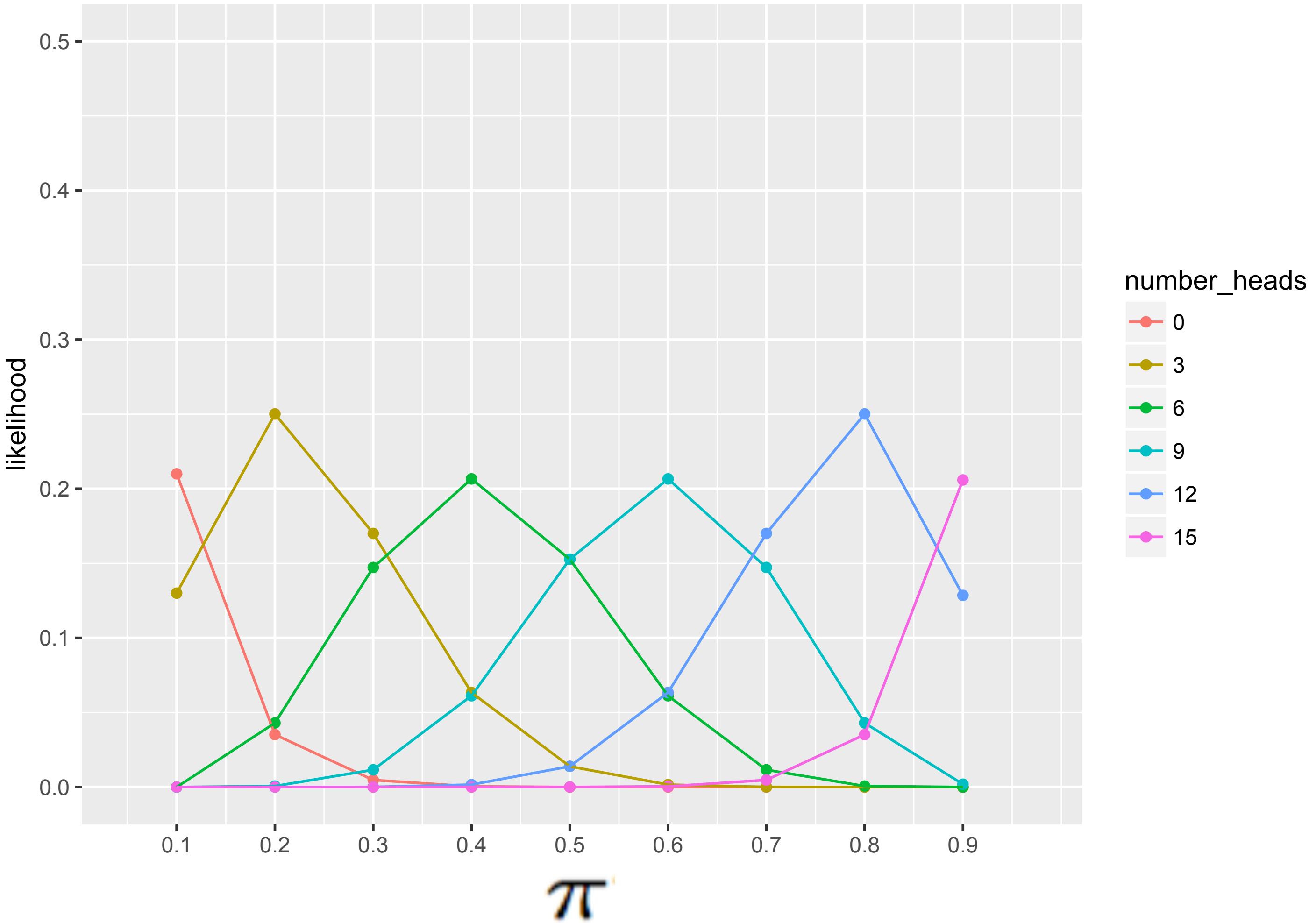
Out of 5 tosses

Maximum Likelihood with 5 coins tosses…



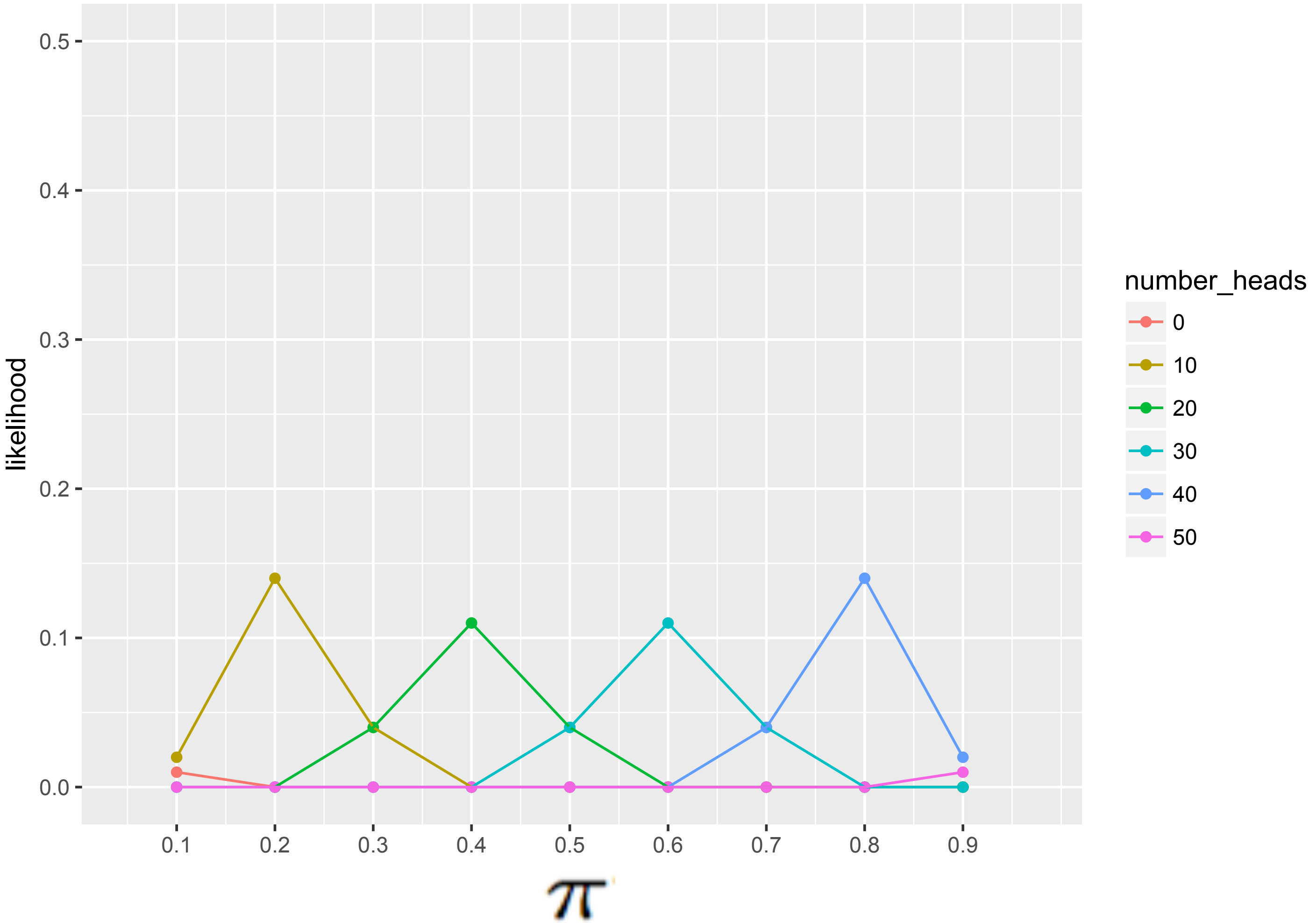
Out of 10 tosses

Maximum Likelihood with 5 coins tosses…



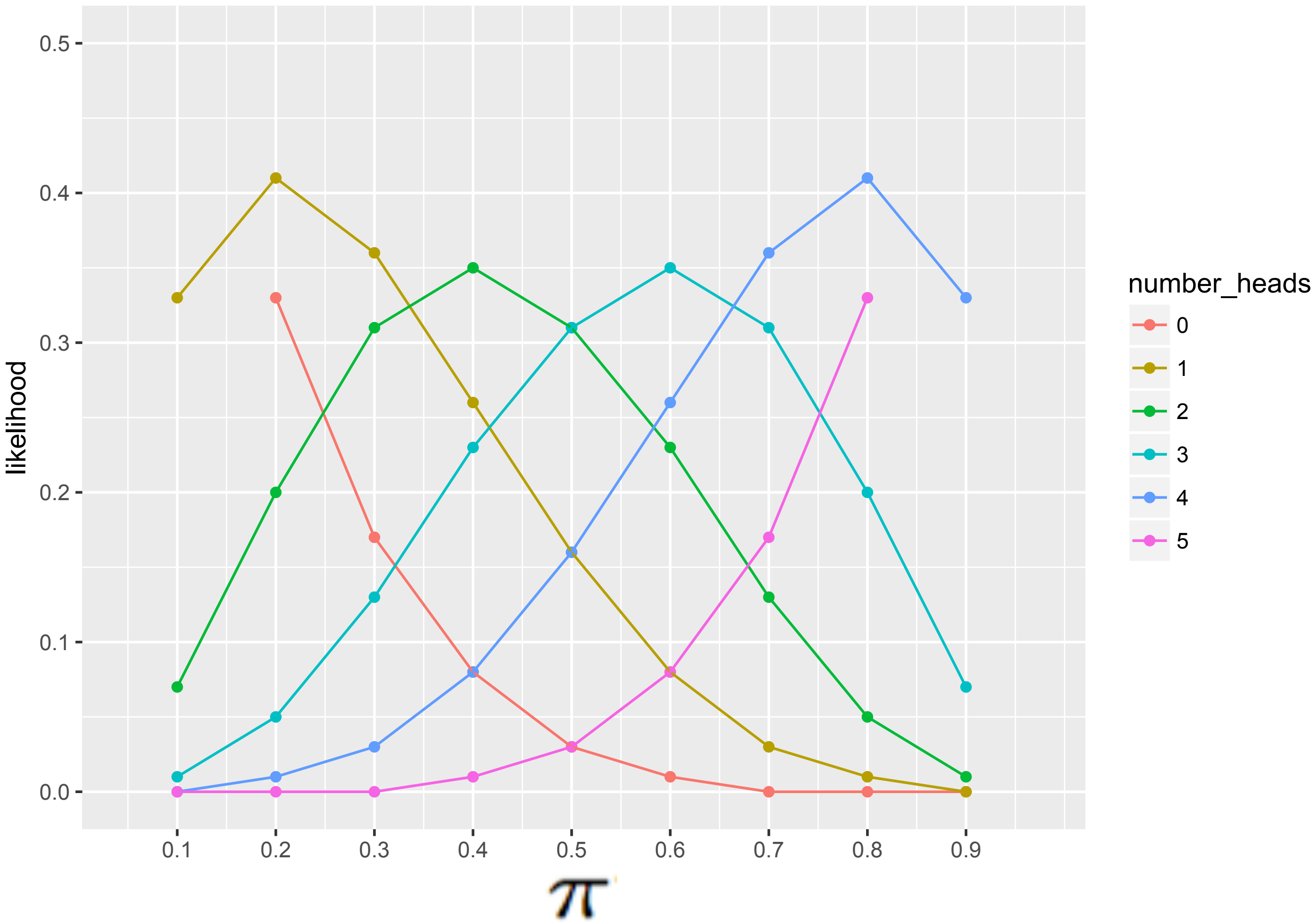
Out of 15 tosses

Maximum Likelihood with 5 coins tosses…



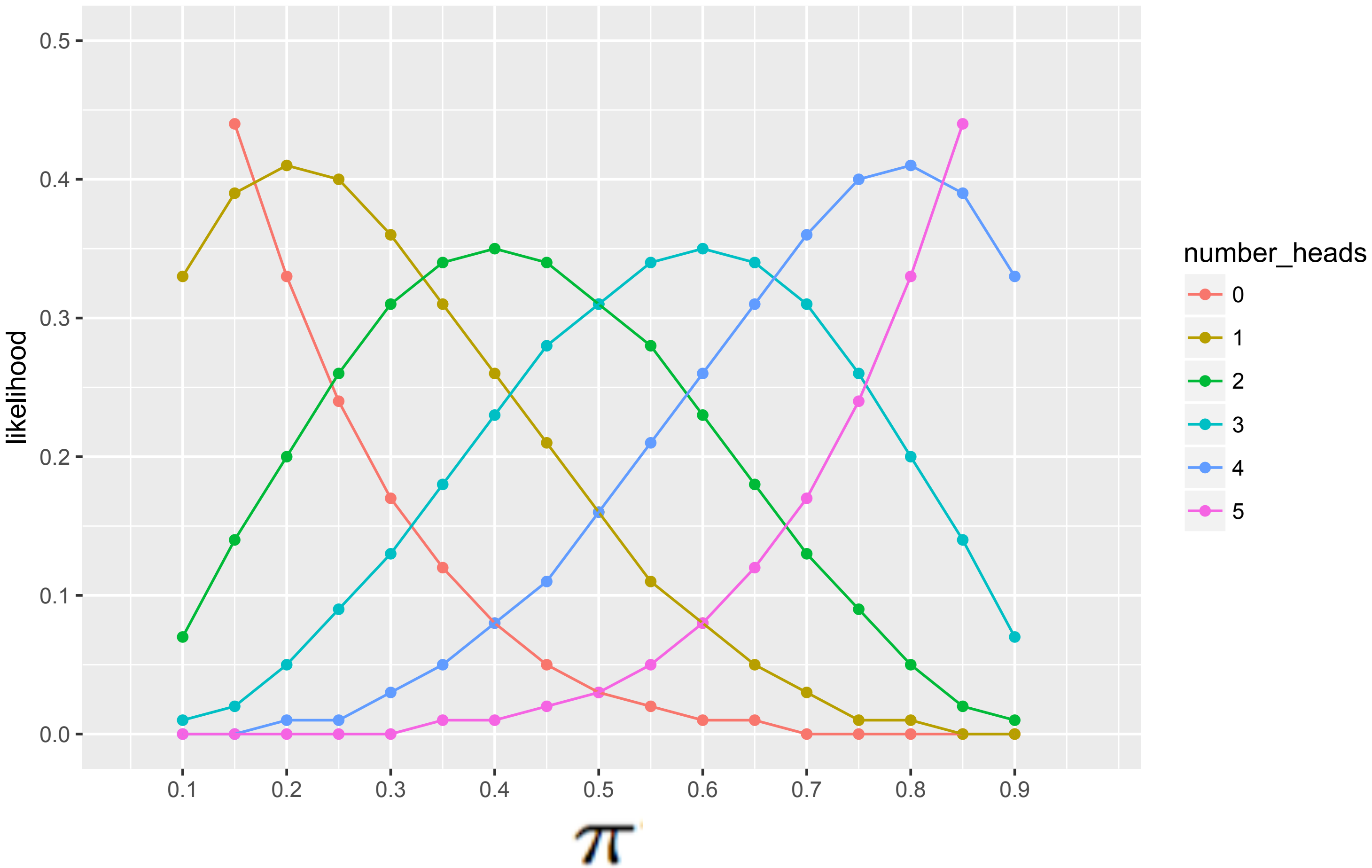
Out of 50 tosses

Maximum Likelihood with 5 coins tosses…

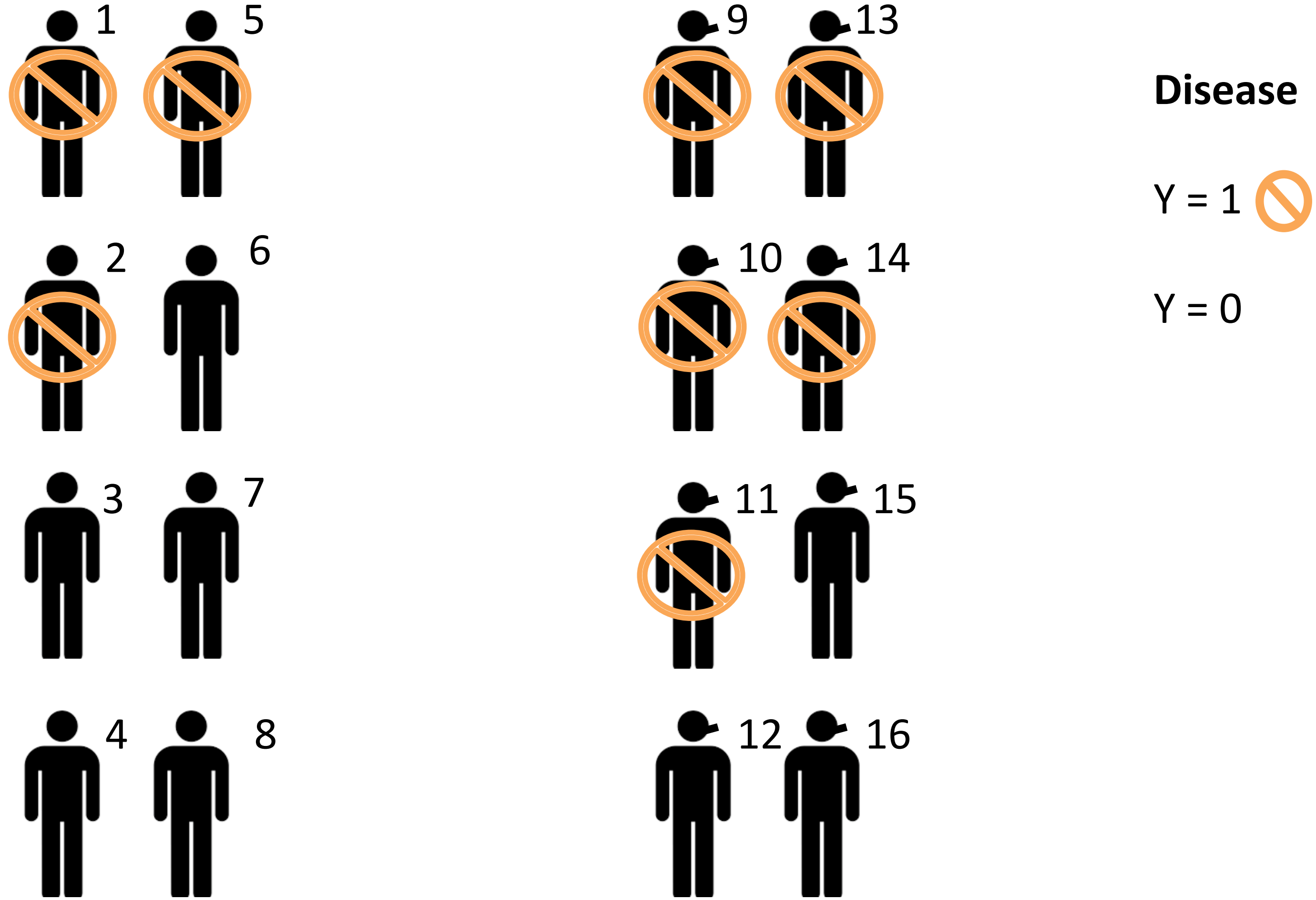


Out of 5 tosses

Maximum Likelihood with 5 coins tosses…



Out of 5 tosses



3/8

non-smokers

X=0

smokers

X=1

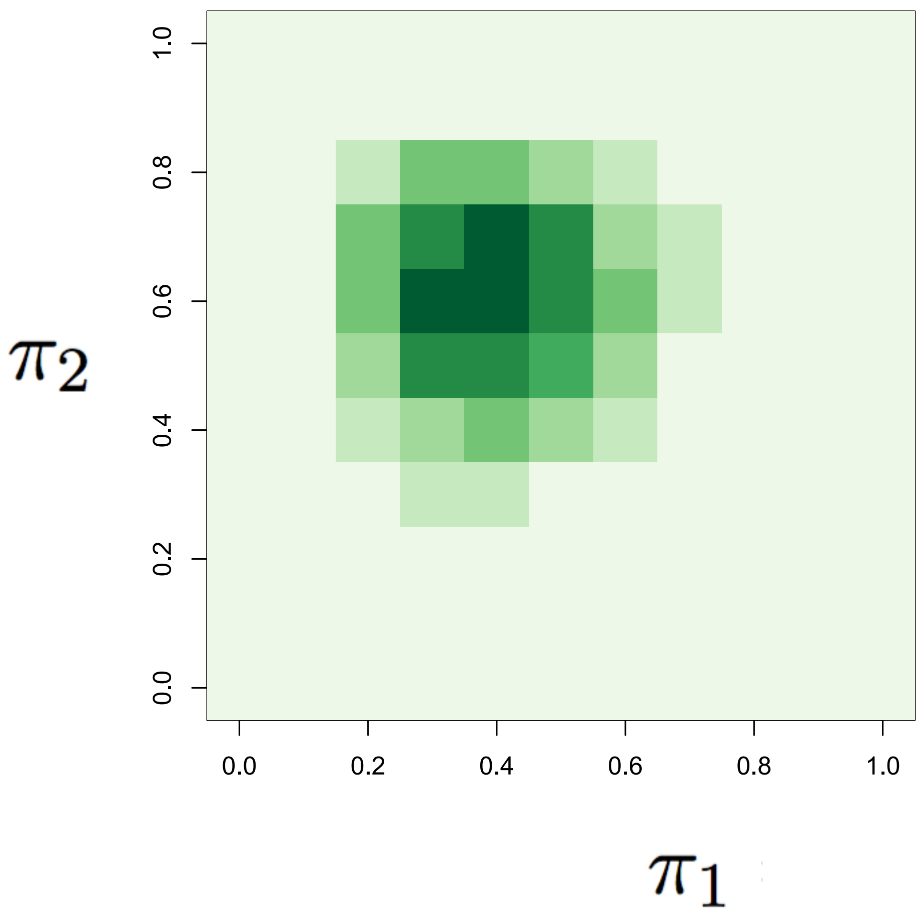
5/8

3/8

non-smokers

5/8

smokers



3/8

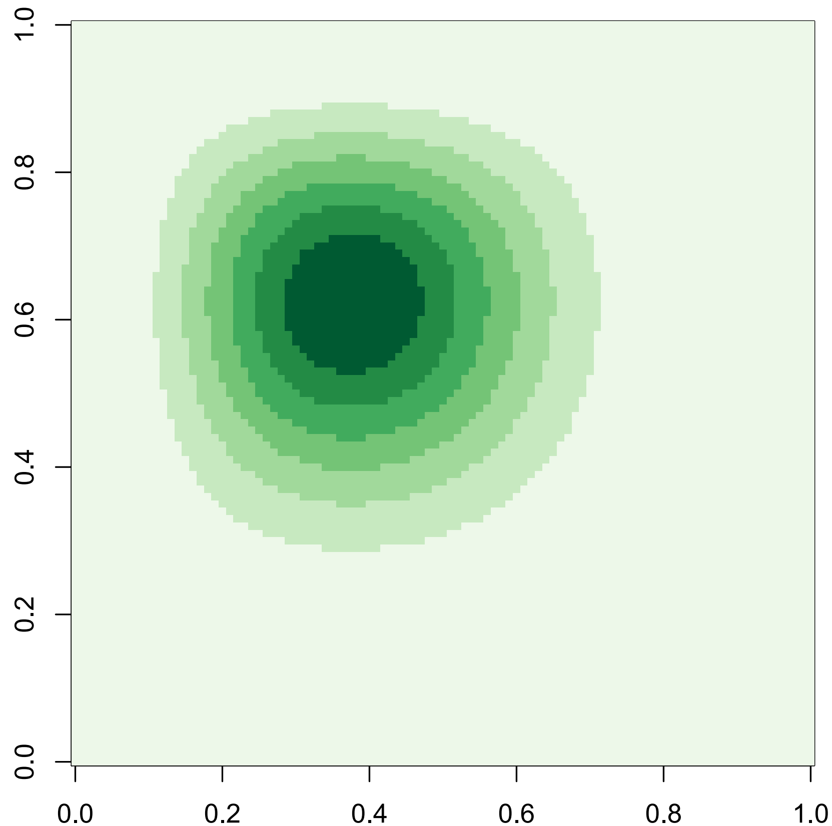
non-smokers

X=0

5/8

smokers

X=1



π_2

π_1

General definition of likelihood

Let data $(y_1, \ldots, y_n)$ be realization of a random vector $(Y_1, \ldots, Y_n)$ with density or model $f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the parameter (usually a vector). The parameter $\boldsymbol{\theta}$ is unknown and the best "guess" is estimated from the data $(y_1, \ldots, y_n)$ by maximizing (over $\boldsymbol{\theta}$) the function:

$$L(\boldsymbol{\theta}; \text{data}) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \boldsymbol{\theta}).$$

After data are observed, consider $(y_1, \ldots, y_n)$ as fixed and $\boldsymbol{\theta}$ as a quantity to be estimated. The parameter value $\boldsymbol{\theta}_1$ is more consistent with the data than $\boldsymbol{\theta}_2$ if

$$L(\boldsymbol{\theta}_1; \text{data}) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \boldsymbol{\theta}_1) > f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \boldsymbol{\theta}_2) = L(\boldsymbol{\theta}_2; \text{data}).$$

General definition of likelihood

Let data $(y_1, \ldots, y_n)$ be realization of a random vector $(Y_1, \ldots, Y_n)$ with density or model $f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the parameter (usually a vector). The parameter $\boldsymbol{\theta}$ is unknown and the best "guess" is estimated from the data $(y_1, \ldots, y_n)$ by maximizing (over $\boldsymbol{\theta}$) the function:

$$L(\boldsymbol{\theta}; \text{data}) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \boldsymbol{\theta}).$$

After data are observed, consider $(y_1, \ldots, y_n)$ as fixed and $\boldsymbol{\theta}$ as a quantity to be estimated. The parameter value $\boldsymbol{\theta}_1$ is more consistent with the data than $\boldsymbol{\theta}_2$ if

$$L(\boldsymbol{\theta}_1; \text{data}) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \boldsymbol{\theta}_1) > f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \boldsymbol{\theta}_2) = L(\boldsymbol{\theta}_2; \text{data}).$$

If $Y_1, \ldots, Y_n$ are independent random variables, then the joint density is a product of univariate densities

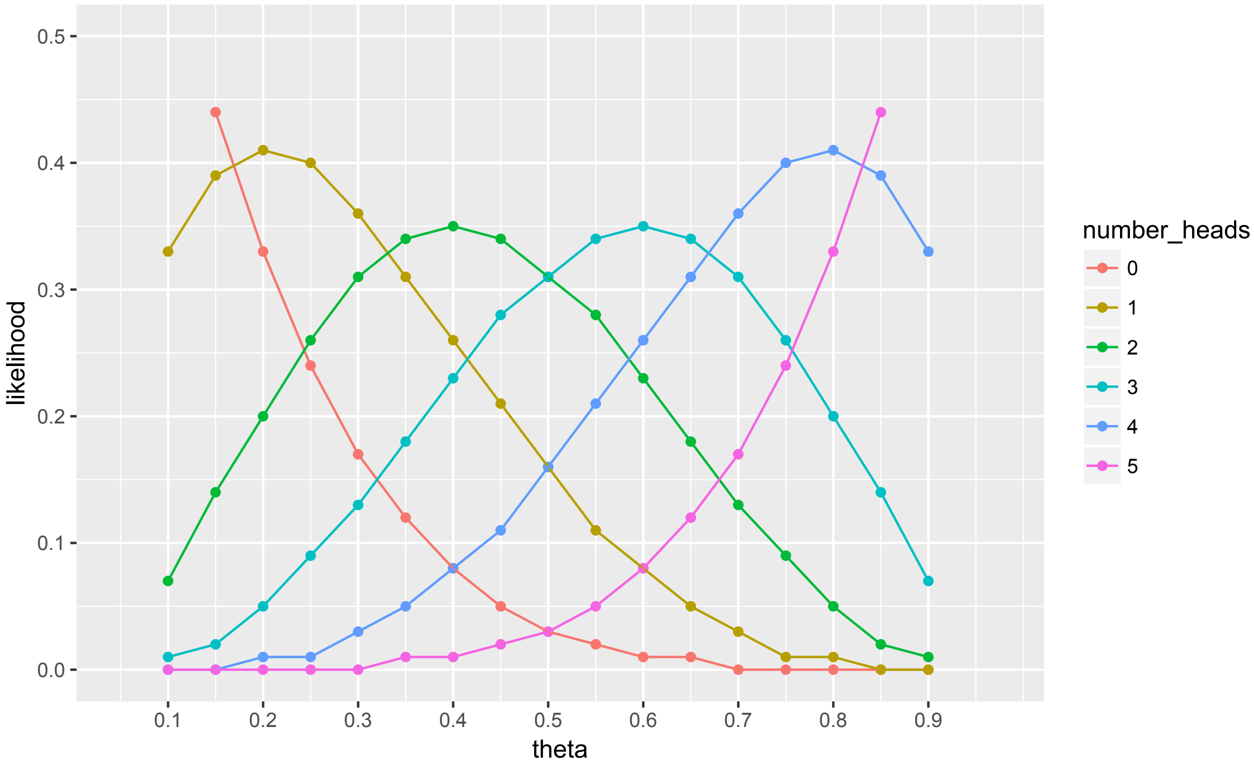
$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_{Y_i}(y_i; \boldsymbol{\theta})$$

The *maximum likelihood estimate* $\hat{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta}; \text{data}))$ in (6.11), or equivalently maximizes $\log L(\boldsymbol{\theta}; \text{data}))$ and minimizes $-\log L(\boldsymbol{\theta}; \text{data})$.

If $Y_1, \ldots, Y_n$ are independent random variables, then the joint density is a product of univariate densities

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_{Y_i}(y_i; \boldsymbol{\theta}). \tag{6.13}$$

Coin toss example:



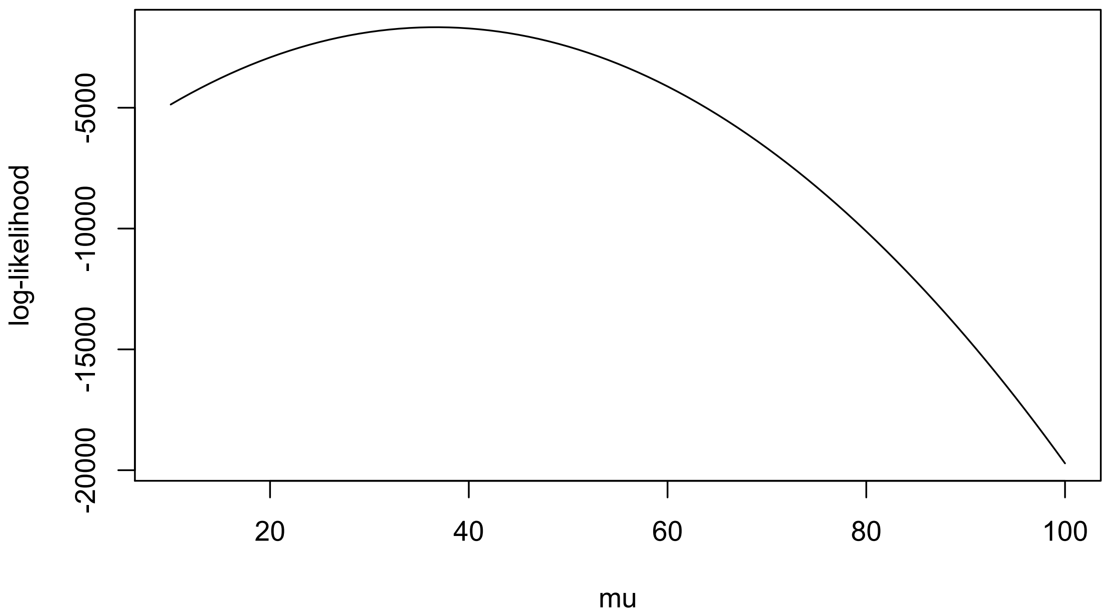
If there are explanatory variables (considered as non-random), and response variables are considered as realizations of random variables, then the likelihood is:

$$L(\boldsymbol{\theta}; \text{data}) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \boldsymbol{\theta}, \text{explanatory } \mathbf{x}_1, \ldots, \mathbf{x}_n). \tag{6.14}$$

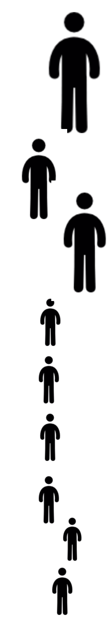
This can get difficult to work with, since it involves lots of multiplications…

So instead we often work with the log-likelihood, log(L).

Consider only sample of Y … what is the maximum likelihood estimate for mu?



Sample, n=9



y

71

54

43

45

21

11

30

45

10

$$Likelihood(\mu|y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp((y_i - \mu)^2/(2\sigma^2))$$

$$logLikelihood(\mu|y) = -\frac{1}{2}n\log(2\pi) - \frac{1}{2}n\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2/(2\sigma^2)$$

Example 6.6 Gaussian regression and homoscedasticity assumption.

$Y_i \sim N(\mu_i = \mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2)$, independently with $\mathbf{x}_i^T = (1, x_{i1}, \ldots, x_{ip})$. Here $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ and $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$. The likelihood and log-likelihood are:

$$L(\boldsymbol{\theta}; \text{data}) = \prod_{i=1}^{n} f_{Y_i}(y_i; \mu_i, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}\sigma} \exp\{-\tfrac{1}{2}(y_i - \mu_i)^2/\sigma^2\} \tag{6.15}$$

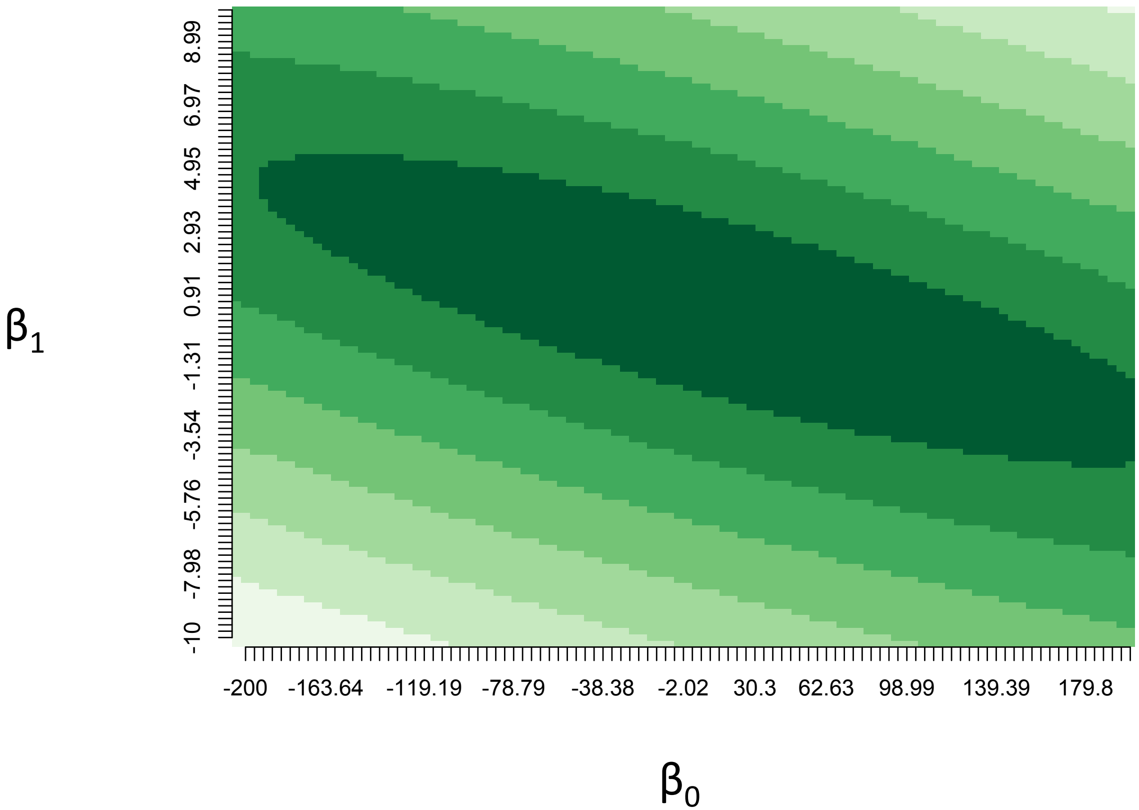
$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}(y_i - \mu_i)^2/\sigma^2\right\}, \tag{6.16}$$

$$\log L(\boldsymbol{\theta}; \text{data}) = -\tfrac{1}{2}n\log(2\pi) - n\log\sigma - \tfrac{1}{2}\sum_{i=1}^{n}(y_i - \mu_i)^2/\sigma^2 \tag{6.17}$$

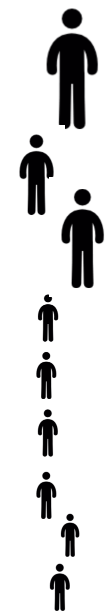
$$= -\tfrac{1}{2}n\log(2\pi) - \tfrac{1}{2}n\log\sigma^2 - \tfrac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2/\sigma^2. \tag{6.18}$$

For any fixed σ^2, maximizing the likelihood is the same as maximizing the log-likelihood, or minimizing $\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$ (that is, least squares). Next, if the parameter σ^2 is also optimized in the likelihood, then its maximum likelihood estimate is $\sum_i e_i^2/n$ instead of $\hat{\sigma}^2 = \sum_i e_i^2/(n-k)$, where e_i are the residuals from least squares. [Check as an exercise].

Now consider Y and X … what is the maximum likelihood estimate for **β**?



Sample, n=9



X	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

$$Likelihood(\beta|y,X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp((y_i - X^T\beta)^2/(2\sigma^2))$$

$$logLikelihood(\beta|y,X) = -\frac{1}{2}n\log(2\pi) - \frac{1}{2}n\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}(y_i - X^T\beta)^2/(2\sigma^2)$$

$\text{Cov}(\hat{\boldsymbol{\theta}})$. Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimate.

Equation for (asymptotic) standard errors, square root of the diagonal of the inverse of the negative Hessian matrix:

$$\left[-\frac{\partial^2 \log L(\boldsymbol{\theta}; \text{data})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\Big|_{\hat{\boldsymbol{\theta}}}\right]^{-1},$$

that is, get the Hessian matrix of negative second order derivatives, take the inverse, extract the diagonal components and take square roots.

The Hessian of g measures the curvature of the negative log-likelihood surface at $\hat{\boldsymbol{\theta}}$. The sharper the curvature is, the smaller the "uncertainty" and the smaller $\pm$ figure for the SE.

The more curved the surface (or parabola if $\boldsymbol{\theta}$ has dimension 1), the larger the Hessian (second derivative) and the smaller the inverse Hessian. SEs come from the sqrt of the diagonal elements of the inverse Hessian.

$\text{Cov}\,(\hat{\boldsymbol{\theta}})$. Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimate.

Equation for (asymptotic) standard errors, square root of the diagonal of the inverse of the negative Hessian matrix:

$$\left[-\frac{\partial^2 \log L(\boldsymbol{\theta};\text{data})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\hat{\boldsymbol{\theta}}}\right]^{-1},$$

that is, get the Hessian matrix of negative second order derivatives, take the inverse, extract the diagonal components and take square roots.
Check what this becomes for the log-likelihood for $Y_i \sim N(\mu_i, \sigma^2)$.

$$\begin{aligned}
-\log L(\boldsymbol{\theta};\text{data}) &= \tfrac{1}{2}n\log(2\pi) + n\log\sigma + \tfrac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2/\sigma^2 \\
&= \tfrac{1}{2}n\log(2\pi) + n\log\sigma + \tfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\sigma^2 \\
-\frac{\partial \log L(\boldsymbol{\theta};\text{data})}{\partial\boldsymbol{\beta}} &= -\frac{\mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} \\
-\frac{\partial^2 \log L(\boldsymbol{\theta};\text{data})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2}
\end{aligned}$$

Inverse is $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. For standard errors, substitute $\hat{\sigma}$ for σ, and get square root of diagonal elements. The above uses matrix/vector derivatives — Section 3.2.

The log-likelihood **for logistic regression**:

$$\log L(\boldsymbol{\beta}; \text{data}) = \sum_{i=1}^{n} (\mathbf{x}_i^T \boldsymbol{\beta}) y_i - \sum_{i=1}^{n} \log[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}]$$

and the

logistic negative log-likelihood

$$-\log L(\boldsymbol{\beta}; \text{data}) = -\sum_{i=1}^{n} (\mathbf{x}_i^T \boldsymbol{\beta}) y_i + \sum_{i=1}^{n} \log[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}]$$

Unfortunately…

There is no closed form solution, but statistical software obtain $\hat{\beta}_0, \hat{\beta}_1$ with an iterative method.

To get standard errors, confidence intervals, we must get the second derivative of **the logistic negative log-likelihood:**

Let $\pi_i = \pi_i(\mathbf{x}_i; \boldsymbol{\beta}) = \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}/[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}] = 1/[1 + \exp\{-\mathbf{x}_i^T \boldsymbol{\beta}\}]$, $1 - \pi_i = 1/[1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}]$. The gradient vector is:

$$-\frac{\partial \log L(\boldsymbol{\beta}; \text{data})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{n} \mathbf{x}_i y_i + \sum_{i=1}^{n} \mathbf{x}_i \frac{\exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}} = -\sum_{i=1}^{n} \mathbf{x}_i y_i + \sum_{i=1}^{n} \mathbf{x}_i \pi_i$$

The Hessian matrix of second order derivatives is:

$$-\frac{\partial^2 \log L(\boldsymbol{\beta}; \text{data})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \pi_i (1 - \pi_i), \tag{1}$$

making use of $\frac{d}{dz} \frac{z}{1+z} = \frac{1}{(1+z)^2}$ and $dz/d\boldsymbol{\beta}^T = \mathbf{x}^T z$, $z = \exp\{\mathbf{x}^T \boldsymbol{\beta}\}$.

When (1) is evaluated at the maximum likelihood estimate, π_i is replaced by $\hat{\pi}_i = 1/[1 + \exp\{-\mathbf{x}_i^T \hat{\boldsymbol{\beta}}\}]$. To check that this is valid, try to get this result in non-matrix form when $p = 1$ (one explanatory variable).

Summary

concept \response type	continuous/normal	binary
linearity	$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$	$\log \frac{\pi_i}{1-\pi_i} = \mathbf{x}_i^T \boldsymbol{\beta}$
no $\mathbf{x}$ effect	$SS(Total) = \sum (y_i - \bar{y})^2$	nulldev$=-2[y_+ \log(\frac{y_+}{n}) + (n - y_+) \log(1 - \frac{y_+}{n})]$
$\mathbf{x}$ effect	$SS(Res; \mathbf{x}) = \sum (y_i - \hat{y}_i)^2$	residdev=-2loglik at MLE
Cov $(\hat{\boldsymbol{\beta}})$	$\hat{\sigma}^2 [\sum_i \mathbf{x}_i \mathbf{x}_i^T]^{-1}$	$[\sum_i \hat{\pi}_i (1 - \hat{\pi}_i) \mathbf{x}_i \mathbf{x}_i^T]^{-1}$
variability explained	R^2, adj R^2	none
-loglik $[\mathbf{X}_{\text{subset}}]$	C_p	AIC = -2(loglik-#parameters)
out-of-sample pred	CVRMSE	out-of-sample misclassification
in-sample	$\hat{\sigma}$	in-sample misclassification

In the above $y_+ = \sum_{i=1}^n y_i$ and the sample proportion is y_+/n.

For AIC=Akaike information criterion, smaller is better.

$C_p = \frac{SS(Res;subset)}{MS(Res:full)} + 2 \times \text{ncol}(\mathbf{X}_{\text{subset}}) - n$; #parameters = ncol$(\mathbf{X}_{\text{subset}})$, ignoring σ^2