DATA MINING 2
Odds and Log Odds

Riccardo Guidotti

a.a. 2021/2022

Contains edited slides from StatQuest
Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:
Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:

Visually, we have 5 games total...
Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:

...1 of which my team will win...
Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:

...and 4 of which my team will lose.
Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:

So the odds are 1...
For example, you might say that the odds in favor of my team winning the game are 1 to 4:
Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:

Alternatively, we can write this as a fraction... \( \frac{1}{4} \)
Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:

Alternatively, we can write this as a fraction... $\frac{1}{4}$

Visually, we have the 1 game my team wins...
Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:

Alternatively, we can write this as a fraction... $\frac{1}{4}$

...divided by the 4 games that my team loses.
Odds Example

For example, you might say that the odds in favor of my team winning the game are 1 to 4:

Alternatively, we can write this as a fraction... \( \frac{1}{4} = 0.25 \) ...if we do the math, we see that the odds are 0.25 that my team will win the game.
Odds Example

Here’s another example: You might say that the odds in favor of my team winning the game are 5 to 3:

Alternatively, we can write this as a fraction... \[ \frac{5}{3} = 1.7 \] ...if we do the math, we see that the odds are 1.7 that my team will win the game.

Note: Odds are not probabilities!!!
Odds vs Probability

The odds are the ratio of something happening (i.e. my team \textit{winning})... to something not happening (i.e. my team \textit{not winning}).
Odds vs Probability

The odds are the ratio of something happening (i.e. my team \textit{winning})...\[ \quad \]
...to something not happening (i.e. my team \textit{not winning}).

Probability is the ratio of something happening (i.e. my team \textit{winning})...\[ \quad \]
...to everything that could happen (i.e. my team \textit{winning and losing}).
Odds vs Probability

In the previous example, the odds in favor of my team **winning** the game are 5 to 3...

\[
\frac{5}{3}
\]

...however, the probability of my team **winning** is the number of games they win (5) divided by the total number of games they play (8)...

\[
\frac{5}{8}
\]

...here's the math...
Odds from Probabilities

\[ \frac{5}{3} = 1.7 \]

In the last example we saw that the odds of winning are 1.7...
Odds from Probabilities

\[ \frac{5}{3} = 1.7 \]

\[ \frac{5}{8} = 0.625 \]

...and the probability of winning is 0.625.
Odds from Probabilities

\[
\frac{5}{3} = 1.7
\]

\[
\frac{5}{8} = 0.625
\]

\[
\frac{3}{8} = 0.375
\]

...the probability of losing is 0.375
Odds from Probabilities

\[ \frac{5}{3} = 1.7 \]

\[ \frac{5}{8} = 0.625 \]

\[ \frac{3}{8} = 0.375 \]

**NOTE:** We could also calculate the probability of **losing** as:

1 - the probability of **winning**
Odds from Probabilities

\[
\text{Odds} = \frac{5}{3} = 1.7
\]

\[
\text{Odds} = \frac{5}{8} = 0.625
\]

\[
\text{Odds} = \frac{3}{8} = 0.375
\]

So, either way, we get the same probability...

1 - the probability of winning = 1 - \(\frac{5}{8}\) = \(\frac{8}{8}\) - \(\frac{5}{8}\) = \(\frac{3}{8}\) = 0.375
Odds from Probabilities

Now let’s take the ratio of the probability of winning to the probability of losing...

The ratio of the probability of winning...

...to the probability of losing
Odds from Probabilities

\[ \frac{5}{3} = 1.7 \]

\[ \frac{5}{8} = 0.625 \]

\[ \frac{3}{8} = 0.375 \]

Alternatively, we can put (1 - the probability of winning) into the denominator...

The ratio of the probability of winning...

...to (1 - the probability of winning)
Odds from Probabilities

\[
\text{odds} = \frac{p}{1-p}
\]

Thus, the ratio of the probabilities ends up being the same thing as the ratio of the raw counts...

The ratio of the probability of winning...

\[
\frac{5}{8} \quad \text{to} \quad \frac{3}{8}
\]

... to (1 - the probability of winning)

\[
\frac{5}{3}
\]

odds = \frac{p}{1-p}
We can see that the worse my team is, the odds of winning get closer and closer to 0.

\[ \frac{1}{32} = 0.031 \]
Log of the Odds

In other words, if the odds are against my team winning, then they will be between 0 and 1.

\[ \frac{1}{32} = 0.031 \]
Log of the Odds

We can see that the better my team is, the odds of winning start at 1 and just go up and up.

\[
\frac{32}{3} = 10.7
\]
Log of the Odds

In other words, if the odds are for my team winning, then they will be between 1 and infinity!

\[
\frac{32}{3} = 10.7
\]
Log of the Odds

Another way to look at this is with a number line...

...to infinity and beyond!
Log of the Odds

The odds of my team losing go from 0 to 1...

...to infinity and beyond!
Log of the Odds

...and the odds of my team **winning** go from 1 to infinity
(and beyond!)

0  1  2  3  4  5  6  ...to infinity
and beyond!
Log of the Odds

The asymmetry makes it difficult to compare the odds for or against my team winning.

...to infinity and beyond!
Log of the Odds

For example if the odds are against 1 to 6, then the odds are $\frac{1}{6} = 0.17...$

...but if the odds are in favor 6 to 1, then the odds are $\frac{6}{1} = 6!$

...to infinity and beyond!
Log of the Odds

Taking the log() of the odds (i.e. log(odds)) solves this problem by making everything symmetrical.
Log of the Odds

For example, if the odds are against 1 to 6, then the log(odds) are:

$$\log(1/6) = \log(0.17) = -1.79$$
Log of the Odds

For example if the odds are against 1 to 6, then the log(odds) are \( \log(1/6) = \log(0.17) = -1.79 \)

...if the odds are in favor 6 to 1, then the log(odds) are \( \log(6/1) = \log(6) = 1.79 \)
Log of the Odds

Using the log function, the distance from the origin (or 0) is the same for 1 to 6 and 6 to 1 odds.
Odds and Log Odds

\[ \frac{5}{3} = 1.7 \]

Earlier we saw that odds can be calculated from counts...
Odds and Log Odds

The ratio of the probability of *winning*... 
...to (1 - the probability of *winning*)

\[
\frac{5}{8} \div \frac{3}{8} = \frac{5/8}{3/8} = \frac{5}{3} = 1.7
\]

...and we saw that the same odds could be calculated from probabilities...
Odds and Log Odds

\[
\text{log(odd)} = \log\left(\frac{5}{3}\right) = \log\left(\frac{p}{1-p}\right) = \log(1.7) = 0.53
\]

...and that means we can calculate the log of the odds with counts or probabilities - either way, we'll get the same value.

The ratio of the probability of winning...

\[
\frac{5}{8} \quad \frac{5}{3} = 1.7
\]

...to \((1 - \text{the probability of winning})\)

\[
\frac{3}{8} \quad \frac{3}{5}
\]
Odds and Log Odds

The log of the ratio of the probabilities is called the **logit function** and forms the basis for logistic regression.

\[
\text{log(odds)} = \log\left(\frac{5}{3}\right) = \log\left(\frac{p}{1-p}\right) = \log(1.7) \approx 0.53
\]
Odds and Log Odds

• Odds are the ratio of something happening to something not happening
• Log odds are the log of the odds
• What’s the big deal?
Odds and Log Odds

To show you what the big deal is all about, if I pick pairs of random numbers that add up to 100 (for example) and use them to calculate the log(odds) and draw a histogram...
Odds and Log Odds

...the histogram is in the shape of a normal distribution!
Odds and Log Odds

This makes the log(odds) useful for solving certain statistics problems - specifically ones where we are trying to determine probabilities about win/lose, or yes/no, or true/false types of situations.
Odds Ratios

When people say “odds ratio”, they are talking about a “ratio of odds”.

[Diagram showing red and blue circles in a pattern]

Odds Ratios
Odds Ratios

When people say “odds ratio”, they are talking about a “ratio of odds”.

So we’ve got a ratio of these odds…

…to these odds.
Odds Ratios

Doing the math gives us...

\[
\frac{2}{4} \div \frac{3}{1}
\]
Just like when we calculate the odds of something, if the denominator is larger than the numerator, the odds ratio will go from 0 to 1...

...to infinity and beyond!
Odds Ratios

...and if the numerator is larger than the denominator, then the odds ratio will go from 1 to infinity (and beyond)...

...to infinity and beyond!
Log of Odds Ratios

...and, just like the odds, taking the log of the odds ratio (i.e. log(odds ratio)) makes things nice and symmetrical.
For example if the odds ratio is \( \frac{2}{4} / \frac{3}{1} \), then the \( \log(\text{odds ratio}) = -1.79 \)
For example if the odds ratio is \( \frac{2}{4}/\frac{3}{1} \), then the log(odds ratio) = -1.79

...and if the odds ratio is \( \frac{3}{1}/\frac{2}{4} \), then the log(odds ratio) = 1.79
Odds Ratios in Action

<table>
<thead>
<tr>
<th>Has the mutated gene</th>
<th>Has Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>23</td>
</tr>
<tr>
<td>No</td>
<td>6</td>
</tr>
</tbody>
</table>
# Odds Ratios in Action

<table>
<thead>
<tr>
<th>Has the mutated gene</th>
<th>Has Cancer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>117</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>210</td>
</tr>
</tbody>
</table>

We can use an “odds ratio” to determine if there is a relationship between the mutated gene and cancer.

If someone has the mutated gene, are the odds higher that they will get cancer?
Odds Ratios in Action

Given that a person has the mutated gene, the odds that they have cancer are...

<table>
<thead>
<tr>
<th>Has Cancer</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>23</td>
<td>117</td>
</tr>
<tr>
<td>No</td>
<td>6</td>
<td>210</td>
</tr>
</tbody>
</table>

\[
\frac{23}{117}
\]
Odds Ratios in Action

<table>
<thead>
<tr>
<th>Has the mutated gene</th>
<th>Has Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes: 23</td>
</tr>
<tr>
<td>No</td>
<td>No: 6</td>
</tr>
</tbody>
</table>

So we'll put that on top of the odds ratio.
Odds Ratios in Action

And given that a person does not have the mutated gene, the odds that they have cancer are....
Odds Ratios in Action

<table>
<thead>
<tr>
<th>Has the mutated gene</th>
<th>Has Cancer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>117</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>210</td>
</tr>
</tbody>
</table>

\[
\text{Odds Ratio} = \left( \frac{23}{117} \right) / \left( \frac{6}{210} \right)
\]

So we'll put that on the bottom of the odds ratio.
Odds Ratios in Action

Here's our odds ratio.

<table>
<thead>
<tr>
<th></th>
<th>Has Cancer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Has the mutated gene</td>
<td>Yes</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>210</td>
</tr>
</tbody>
</table>

\[
\text{Odds Ratio} = \frac{23}{117} \times \frac{210}{6} = \frac{23 \times 210}{117 \times 6}
\]
Odds Ratios in Action

<table>
<thead>
<tr>
<th>Has the mutated gene</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>23</td>
<td>117</td>
</tr>
<tr>
<td>No</td>
<td>6</td>
<td>210</td>
</tr>
</tbody>
</table>

We do the math...

\[
\frac{23}{117} \div \frac{6}{210} = \frac{0.2}{0.03} = 6.88
\]
Odds Ratios in Action

<table>
<thead>
<tr>
<th>Has the mutated gene</th>
<th>Has Cancer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>23</td>
<td>117</td>
</tr>
<tr>
<td>No</td>
<td>6</td>
<td>210</td>
</tr>
</tbody>
</table>

...and the odds ratio tells us that the odds are 6.88 times greater that someone with the mutated gene will also have cancer.

$$\frac{23}{117} \div \frac{6}{210} = \frac{0.2}{0.03} = 6.88$$
Odds Ratios in Action

<table>
<thead>
<tr>
<th>Has the mutated gene</th>
<th>Has Cancer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>23</td>
<td>117</td>
</tr>
<tr>
<td>No</td>
<td>6</td>
<td>210</td>
</tr>
</tbody>
</table>

\[
\frac{23}{117} = \frac{0.2}{0.03} = 6.88
\]

\[
\log(6.88) = 1.93
\]

...and the log(odds ratio) is 1.93.
Odds Ratios in Action

<table>
<thead>
<tr>
<th>Has the mutated gene</th>
<th>Has Cancer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>23</td>
<td>117</td>
</tr>
<tr>
<td>No</td>
<td>6</td>
<td>210</td>
</tr>
</tbody>
</table>

What does all this mean?

\[
\frac{23}{117} = \frac{0.2}{0.03} = 6.88
\]

\[\log(6.88) = 1.93\]
Odds Ratios in Action

...larger values mean that the mutated gene is a good predictor of cancer. Smaller values mean that the mutated gene is not a good predictor of cancer.

<table>
<thead>
<tr>
<th>Has the mutated gene</th>
<th>Has Cancer</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>23</td>
<td>117</td>
</tr>
<tr>
<td>No</td>
<td>6</td>
<td>210</td>
</tr>
</tbody>
</table>

\[
\frac{23}{117} \div \frac{6}{210} = \frac{0.2}{0.03} = 6.88
\]

\[
\text{log}(6.88) = 1.93
\]
Odds Again

• Given some event with probability $p$ of being 1, the odds of that event are given by:

  $$\text{odds} = \frac{p}{1-p}$$

• When we go from Normal to High, the odds of being Sick triple:
  - Odds ratio: $0.293/0.111 = 2.64$
  - 2.64 times more likely to be Sick with high values

<table>
<thead>
<tr>
<th>Value</th>
<th>Sick</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
</tr>
<tr>
<td>Normal</td>
<td>402</td>
<td>3614</td>
<td>4016</td>
</tr>
<tr>
<td>High</td>
<td>101</td>
<td>345</td>
<td>446</td>
</tr>
<tr>
<td>Total</td>
<td>503</td>
<td>3959</td>
<td>4462</td>
</tr>
</tbody>
</table>

The odds of being sick if you have a Normal value are:
- $\text{Odds(Sick|Normal)} = \frac{P(\text{sick})}{1-P(\text{sick})} = \frac{402/4016}{1-(402/4016)} = 0.1001 / 0.8889 = 0.111$

The odds of being not sick with a Normal value is the reciprocal:
- $\text{Odds(not Sick|Normal)} = 0.8999/0.1001 = 8.99$

For the High value we have
- $\text{Odds(Sick|High)} = 101/345 = 0.293$
- $\text{Odds(not Sick|High)} = 345/101 = 3.416$
Logit Transform

- The logit is the natural log of the odds
- \text{logit}(p) = \ln(\text{odds}) = \ln\left(\frac{p}{1 - p}\right)
• Regression. Appendix D. Introduction to Data Mining.