AN INTRODUCTION TO BAYESIAN ANALYSIS AND BAYESIAN REGRESSION IN R

Michael Crawford & Michael Cooney

Barnett Waddingham
BAYESIAN STATISTICS IN 5 MINUTES
WHICH IS IMPOSSIBLE
SO LET'S RATHER TRY TO...
LEARN ABOUT BAYES RULE IN 5 MINUTES
SO WHAT IS **BAYES RULE?**
\[ P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)} \]
HOW DO WE USE THIS?
YOU USE **BAYES RULE** ALL THE TIME
YOU JUST DON'T KNOW IT
BAYES RULE

SHOWS HOW YOUR BELIEFS CHANGE

WHEN YOU GET NEW INFORMATION
LET'S START WITH SOMETHING FAMILIAR
TOSSING A COIN
LET'S TOSS THE COIN A FEW TIMES

AFTER EACH TOSS

TELL ME IF THE COIN IS FAIR
AT OUTSET IT'S REASONABLE TO ASSUME THE COIN IS FAIR THIS IS YOUR INITIAL BELIEF
<table>
<thead>
<tr>
<th>Toss</th>
<th>Fair? Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>First toss: H</td>
<td>Yes</td>
</tr>
<tr>
<td>Second toss: H</td>
<td>Yes</td>
</tr>
<tr>
<td>Third toss: H</td>
<td>Yes</td>
</tr>
<tr>
<td>Fourth toss: H</td>
<td>Yes doubts set in</td>
</tr>
<tr>
<td>Fifth toss: H</td>
<td>Possibly</td>
</tr>
<tr>
<td>Sixth - Tenth toss: H</td>
<td>Probably not</td>
</tr>
<tr>
<td>Eleventh - Twentieth toss: H</td>
<td>Definitely not</td>
</tr>
</tbody>
</table>

Chances of this are: 1 in 1,048,576
YOU HAVE JUST INTUITIVELY APPLIED BAYES RULE WITHOUT REALISING IT
AS YOU GOT MORE INFORMATION

YOU ALTERED YOUR BELIEF

THAT THE COIN IS FAIR
NOW LET'S LOOK AT BAYES RULE AGAIN
BAYES RULE ALLOWS YOU QUANTIFY THIS QUALITATIVE PROCESS
\[ P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \]
WE CAN SIMPLIFY THIS
\[ P(A \mid B) \propto P(B \mid A) P(A) \]
$P(A)$

IS THE PRIOR DISTRIBUTION
AND REPRESENTS OUR INITIAL BELIEF
$P(B \mid A)$

is the likelihood model

and updates as the data arrives
$P(A \mid B)$

is the posterior distribution

our updated beliefs for $A$
Prior Distribution Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>10</td>
</tr>
<tr>
<td>Beta</td>
<td>10</td>
</tr>
</tbody>
</table>

Coin Toss Results

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heads</td>
<td>20</td>
</tr>
<tr>
<td>Tails</td>
<td>0</td>
</tr>
</tbody>
</table>

Calculate  Reset
USING THESE SIMPLE BUILDING BLOCKS

WE CAN BUILD POWERFUL MODELS
WHY BAYESIAN?
INTERPRETABILITY

Natural interpretation of output

Credibility region vs Confidence interval
SPARSE DATA PROBLEMS

Claims reserving

Pricing

Account segmentation
CLAIMS RESERVING

Changes in reserving philosophy

Assessing multiple insurers
THE BAYESIAN APPROACH
POSTERIOR DISTRIBUTION

\[ p(\theta | D) = \frac{\int p(D | \theta) p(\theta)}{\int p(D)} \]

\( p(\theta) = \) prior distribution of \( \theta \)
\( p(\theta | D) = \) posterior distribution of \( \theta \) given \( D \)
p(θ | D) \propto \int p(D | θ) p(θ)
How do we calculate this integral?
STAN

MCMC via HMC

Probabilistic Programming Language

C++ backend

Excellent online community
Why bother?
Captures uncertainty

Easy to iterate and improve

 Allows generative modelling

Hierarchical modelling for sparse data
PITFALLS

Learning ‘cliff’

Requires aspects of physics, computation, statistics

Can seem overwhelming
Start with simple linear model
LINEAR MODELS
GETTING STARTED

Ordinary Least Squares (OLS)

Input variables $X$, parameters $\beta$

$$y = \beta X + \epsilon,$$
$$\epsilon \sim \mathcal{N}(0, \sigma)$$

Constant variance $\sigma$. 
Rethink linear models in Bayesian language

Need probability model
BASIC ASSUMPTIONS

Data distributed as Normal

Mean for each point is linear function of $X, \beta X$

\[ y \sim \mathcal{N}(\beta X, \sigma) \]
### SIMPLE CLAIMS MODEL

<table>
<thead>
<tr>
<th>log_loss</th>
<th>lawyer</th>
<th>gender</th>
<th>seatbelt</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.553632</td>
<td>yes</td>
<td>male</td>
<td>yes</td>
<td>50</td>
</tr>
<tr>
<td>2.388029</td>
<td>no</td>
<td>female</td>
<td>yes</td>
<td>28</td>
</tr>
<tr>
<td>-1.108663</td>
<td>no</td>
<td>male</td>
<td>yes</td>
<td>5</td>
</tr>
<tr>
<td>2.401253</td>
<td>yes</td>
<td>male</td>
<td>no</td>
<td>32</td>
</tr>
<tr>
<td>-1.980502</td>
<td>no</td>
<td>male</td>
<td>yes</td>
<td>30</td>
</tr>
<tr>
<td>-1.174414</td>
<td>yes</td>
<td>female</td>
<td>yes</td>
<td>35</td>
</tr>
<tr>
<td>1.263562</td>
<td>yes</td>
<td>male</td>
<td>yes</td>
<td>19</td>
</tr>
</tbody>
</table>

\[
\text{log(loss) } \sim \text{ lawyer} + \text{ seatbelt} + \text{ gender} + \text{ age}
\]

‘Formula notation’
MLE MODEL (IN R)

model_lm <- lm(log_loss ~ lawyer + seatbelt + gender + age,
               data = modeldata_tbl)
RSTANARM PACKAGE

Pre-built models

Linear models, GLMs, ANOVA, etc.

Built for ease of use
RSTANARM VERSION

```r
model_stanlm <- stan_lm(log_loss ~ lawyer + seatbelt + gender + age
  ,prior = R2(location = 0.8)
  ,data = modeldata_tbl)
```
SUMMARY

Bayesian output captures uncertainty

More and more common

Learning curve
FURTHER READING

Stan Documentation/Vignettes/Case Studies, Stan Core Team et al.
http://www.mc-stan.org

Data Analysis Using Regression and Multilevel/Hierarchical Models, Gelman and Hill
http://www.stat.columbia.edu/~gelman/arm/

Statistical Rethinking, McElreath
http://xcelab.net/rm/statistical-rethinking/

Doing Bayesian Data Analysis, Kruschke
https://sites.google.com/site/doingbayesiandataanalysis/
QUESTIONS?

Nah, we're running outta time. Seriously.

michael.crawford@barnett-waddingham.co.uk

michael.cooney@barnett-waddingham.co.uk

THANKS FOR LISTENING

The R code is available on request