OUTLINE

1. Basics
2. Binomial models
3. Continuous variable models

INTRODUCTION TO BAYESIAN REASONING FOR BEGINNERS *

BY WENDY OLSEN

2021

CONTACT WENDY.OLSEN@MANCHESTER.AC.UK AT THE CATHIE MARSH CENTRE, UNIVERSITY OF MANCHESTER

* AIMING AT SOCIOLOGY AND ECONOMICS AND RELATED AREAS OF RESEARCH

ACKNOWLEDGEMENTS

Thank you to Arek Wiśniowski

Social Statistics

University of Manchester

This unit is linked to the MSc In Social Research Methods & Statistics and the

PhD programme in Social Statistics, School of Social Sciences, Univ of Manchester.

See https://www.manchester.ac.uk/study/masters/courses/list/06097/msc-social-research-methods-and-statistics/ and https://www.socialsciences.manchester.ac.uk/social-statistics/

3 BASIC CONCEPTS, PROBABILITY CAN BE SCALED. JOINT AND CONDITIONAL PROBABILITY.

- Conditional probability

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

- Bayes theorem (1763)

$$P(D|B)=\frac{P(B|D)P(D)}{P(B)}$$

4 POSTERIOR PROBABILITY

- **Summary: Bayes' Theorem says that**
- $\mathbf{P(\theta|D)} = \frac{P(\mathbf{D}|\theta)\cdot P(\theta)}{P(\mathbf{D})}$
- $\mathbf{P(\theta|D) \propto P(D|\theta)\cdot P(\theta)}$
- **And** $\mathbf{P(\theta|reality)} \leftarrow\rightarrow \mathbf{P}(\hat{\theta}|D)$
- **And** $\mathbf{P}(\hat{\theta}|D) \propto$ **Likelihood · Prior**
- Restating these in words depends on the concept of 'posterior' which means afterward.

5 POSTERIOR PROBABILITY

- **The Posterior function depends on the data $P(\theta|\mathbf{D})$**
- **$P(\theta|\mathbf{D}) = \frac{P(\mathbf{D}|\theta)\cdot P(\theta)}{P(\mathbf{D})}$**
- **In turn, the posterior is proportional to the likelihood and the prior (product).**
    - **$P(\theta|\mathbf{D}) \propto P(\mathbf{D}|\theta)\cdot P(\theta)$**
- **I like to say it this way: $P(\theta|\text{reality}) \leftarrow\rightarrow P(\hat{\theta}|D)$ so D is our Information.**
- **And in turn, we estimate this by multiplying the Likelihood · Prior.**
- Understanding these words depends on the concept of ‘likelihood’ which is a function.

6 LIKELIHOOD FUNCTION

- concept of ‘likelihood’ which is a function.

$$P(\text{Model} \mid \text{Data}) \propto P(\text{Model}) * \boxed{P(\text{Data} \mid \text{Model})}$$

It is easy to derive this rule from the definitions of conditional probability and joint probability.

Also note, the red item is the likelihood function!

REVIEW OF LIKELIHOODS

- First the probability of one data item, such as Yi, depends on the distribution from which it is drawn. For the sake of Bayes theorem we are going to insert an assertion about this distribution type into the equation, creating a likelihood.
-

P(Datum|Model)

For a whole data set, i= 1…n, we have to use the product operator to get the overall probability as a number, or as a function of the underlying parameters.

P(Data|Model) = Π
(

KEY SOURCES

- Basic: Michael J. Crawley, 2015, *Statistics: An Introduction Using R.* London: Wiley. 2nd ed.
- Casella, G., and R.L. Berger, 1990, *Statistical Inference*, Wadsworth, to review maximum likelihood, PDFs, CDFs and various tests.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, 2013, *Bayesian Data Analysis,* 3rd ed., London: CRC Press and Taylor & Francis, Chapman and Hall. Series: Texts in Statistical Science.
- Gelman, A., 2004, "Parameterization and Bayesian Modeling", *Journal of the American Statistical Association*, 99 537-545.
- Crawley, M.J., 2013, *The R Book*, 2nd ed., London: Wiley. (Maximum likelihood is covered in Chs 7 and 9, and regression ch 10, Bayesian statistics Ch 22, including BUGS and JAGS in R)

9 HOW WE PROGRAMME THIS IN R R2JAGS OR STAN OR OTHER

- **Here is a likelihood function, seen in the WinBUGS code format:**
- ### LIKELIHOOD ###

```
for (j in 1: N.obs){

     for (i in 1: N.X){

          X.row[i, j] <- X.Eff[i, X[j, i]]

     }

     for (i in 1: N.Z){

          Z.row[i, j] <- Z.Eff[i, Z[j, i]]

     }

     log(mu[j]) <- Beta0 + log(Offset[j]) + sum(X.row[, j]) + sum(Z.row[, j])

     Y[j] ~ dpois(mu[j])

} }
```

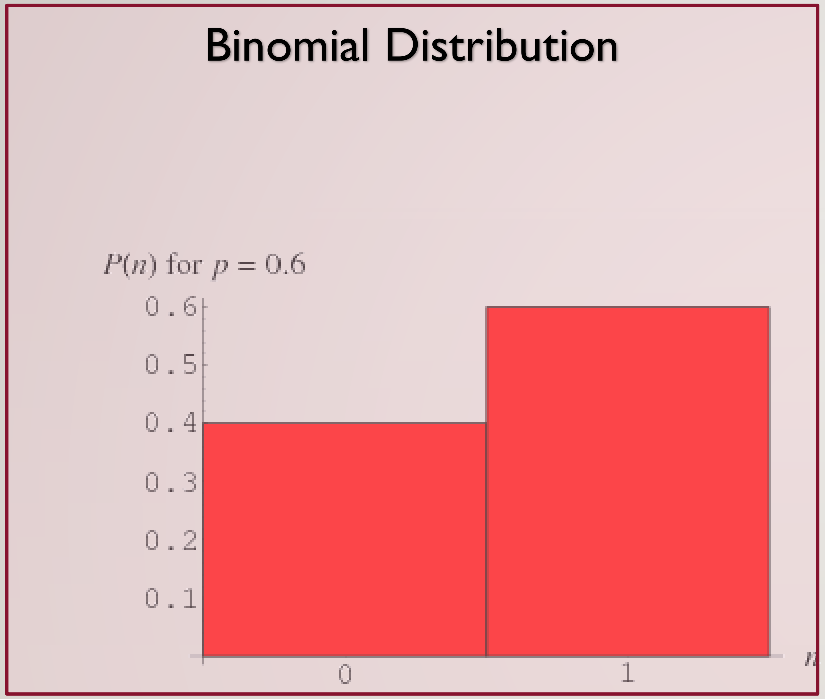
This code represents a Poisson Model.

TYPES OF BAYESIAN MODELS FROM A PRACTITIONER POINT OF VIEW

OUTLINE

1. Basics
2. Binomial models
3. Continuous variable models

- As you are learning, the first example is usually a coin toss or fail/succeed of a treatment T. (MODEL TYPE 1)
- T=0 or T=1 and this is modelled as a single Bernoulli Trial.
- The resulting distribution is the probability density function (PDF), and it has a binomial distribution shape.
- The cumulative distribution function is very simple. From this we can read off a probability for a given level of probability of T.



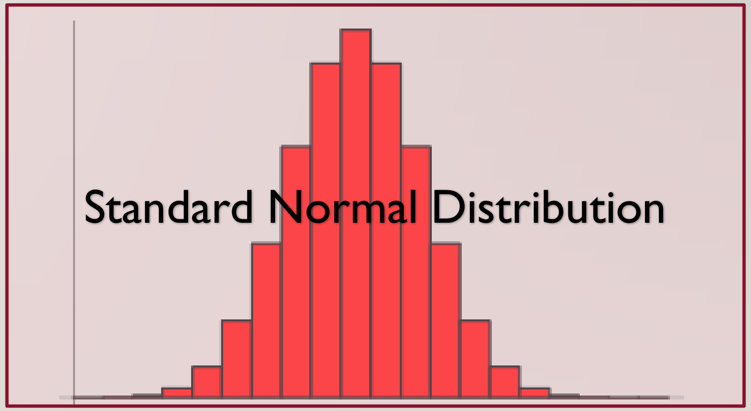
Binomial Distribution

$$X \sim \text{Binomial}(n,p)$$

Parameters n and p

$$E(X) = np$$

$$Var(X) = np(1-p)$$



Standard Normal Distribution

The pdf above does not look like the standard normal pdf.

In Bayesian reasoning we will use up to 12 or 14 distribution types.
Each has a pdf and a cdf. PDF=Probability distribution function.
CDF = the cumulative distribution function.

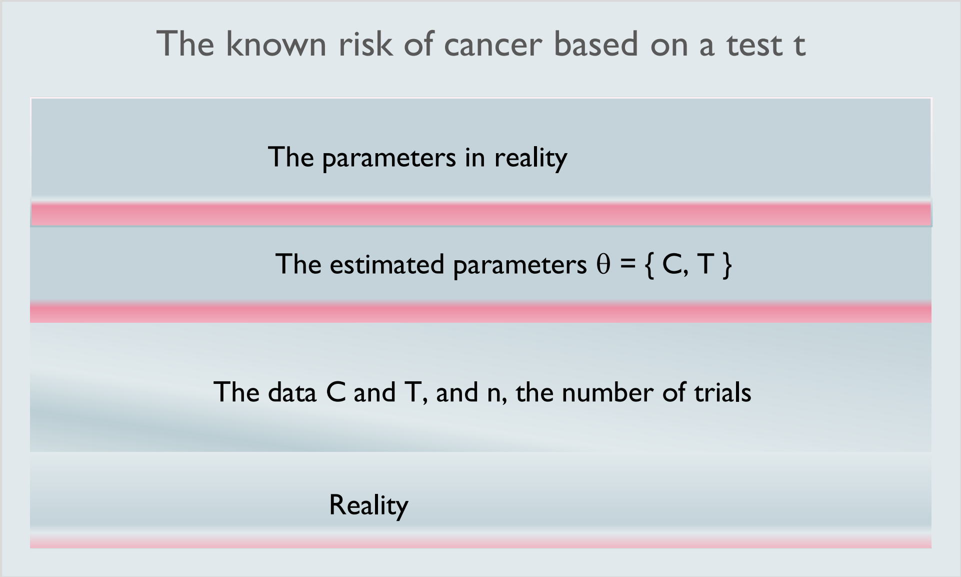
A HIERARCHICAL MODEL OF A CANCER TEST RESULT

- Cancer test t = 0 or t = 1

- If you know a parameter, put it into your model.

- This would be an informed prior

The binomial function is funny in having just one parameter p, plus the number of trials, n



The known risk of cancer based on a test t

The parameters in reality

The estimated parameters θ = { C, T }

The data C and T, and n, the number of trials

Reality

Terminology for Exercise 1:

Cowles (2013):
Sensitivity is 0.89 in Bauer (1997).
The probability of a false negative is 1-sensitivity of the test.

And
Specificity is 0.94 in Bauer (1997).
The probability of a false positive is 1-specificity.

See Exercise 1 and the document containing the results for cancer tests according to different levels of sensitivity of a mammogram screening test.

Yudkowsky.net/rational/bayes/

14 ESSENTIAL DIFFERENCES

Example: in a study of pay gaps, Model 1 assumes that the wage is normally distributed, and Model 2 assumes it is gamma distributed. Model 3 allows wage as a sum of mini-distributions.

BAYESIAN METHODS

- Estimates the probability of a model
- For 2 models, it can tell which is more probable
  - (BAYES FACTOR) [given some data]
- Does not assume the model is correct, so after doing 4 models you can average these

THE CLASSICAL HYPOTHESIS TESTING REGIME

- Assumes error terms are normally distributed
- Assumes no correlation of the X's other than what is modelled
- Assumes the correct model
- Tests mainly the parameters

15 MORE ON THE BAYES FACTOR APPROACH

This is an optional approach.

BAYESIAN METHODS

- First calculate the probability of Model 1 given Y, and then the probability of Model 2 given Y.
- For 2 models, we can tell which is more probable
  - (BAYES FACTOR | given some data)
- See Cowles, Ch. 11, pg 209

THE CONCEPT OF MODEL AVERAGING

- A Bayesian does not assume the model is correct, so after doing 4 models you can average these
- I am unsure this is a good idea, as you have "posited" all four models.
- Delphi method is a better idea.

EXPLAINING WHY THE BAYES FACTOR FORMULA WORKS SO WELL

PRIOR ODDS VS POSTERIOR ODDS

- $\frac{\Pr(\mathrm{M}_1)}{\Pr(\mathrm{M}_0)} = \frac{\Pr(\mathrm{H}_1)}{\Pr(\mathrm{H}_0)}$ this is the prior odds of model 1 being the true state of affairs in the world.
- We have assumed there are only 2 possible states of affairs in the world.

THE POSTERIOR ODDS ARE:

- $\frac{\Pr(\mathrm{M}_1|y)}{\Pr(\mathrm{M}_0|y)}$
- Looking at this, it has elements that we know, insofar as we know the data y and the distribution formats, range restrictions and relationships.
- Chapters 1 and 11 of Cowles (2013) contain the mammogram example using Bauer (1997) data

17 SO WHAT ARE P VALUES?

BAYESIAN

- To them, a P-value is a mythical beast, a simplistic social construction.
- I advise we interpret Bayesian approach from a realist meta approach.

CLASSICAL REASONING

- P value is the probability of being wrong
- The confidence level is the probability of getting this answer, in this 95% range, over many multiple hypothetical repeat samples with replacement from the population
- Thus it's focused on the likelihood of the data, given that the model is perfectly right
- Or… given that it's the true model.

We got used to these conditional statements. We can get used to the other.

SECOND TYPE OF MODEL: A HIERARCHICAL MODEL OF A CONTINUOUS OUTCOME

- Now we can model the probability T as a parameter, which is a continuous random variable.
- Its distribution has a mean, mu of T. We estimate mu by taking E(pdf(T)).
- This Probability Density Function (distribution) also has a standard deviation. Please note, this is

not a standard error.

  - In Bayesian statistics we move away from the concept of standard error.
  - The square root of the Var(PDF(T)) is known as the standard deviation
  - The algebra of decomposing the variance has to be familiar (Casella and Berger)
- We can estimate this standard deviation. Our notation for it will be Tau of PDF(T).
- The standard deviation of the probability estimator is a measure of the width of the random variable PDF(T)'s distribution. The new RV, mu, is not the actual value T and it does not have to have a symmetric distribution. Of course it's not necessarily a normal distribution. (See red diagrams)

NEARLY DONE

1. Basics
2. Binomial models
3. Continuous variable models

Kass and Raftery 1995

A DIAGRAM FOR A HIERARCHICAL MODEL

- ‘doodle’ in WinBUGS (BUGS means …Gibbs Sampler…which does MCMC sampling…which is a form of simulation to estimate a model)
- Outermost layer= hyperpriors
- Inner layer = two types of element.
  - The Data
  - The Parameters
  - These in turn will lie in Directed Acyclic Graphs, or in order of causality.

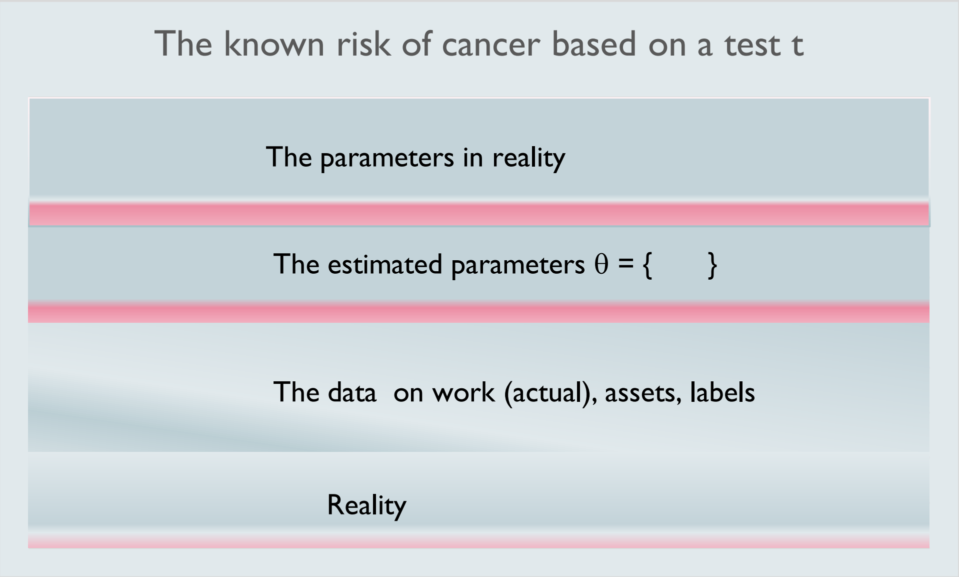
A HIERARCHICAL MODEL OF A LABOUR OUTCOME WITH FORMAL & INFORMAL MARKETS

- Labour which is paid no, W = 0 or yes, W = 1
- Second part of model: Which sector they work in.
  - Multinomial logit. A risk for each sector j.
  - Third part is gender, Sex=0 (male) or 1 (female)
- If you know a parameter, put it into your model.
- You may use an

informed prior

Or uninformed flat prior

The logit function can be found in Generalised Linear Models or in Bayesian models.



The known risk of cancer based on a test t

The parameters in reality

The estimated parameters θ = {   }

The data on work (actual), assets, labels

Reality

The multinomial logit function combined with Logit would be a structural equation model in Classical Statistical reasoning.

BUZZ TASK

- Suppose we want to model labour supply. We know there is measurement error in the measure of Yes/No work of women (here, work is the paid work women do)
- Using words, build a short causal model of the risk of working, given the risk of measuring it wrongly and the reasons why women work. Use a diagram if you wish.

Hint: Why do we care? Because we could use the time-use survey or local data on women's real-life decisions to build up estimates for our model. These would be **informative.**

ANSWER TO THIS TASK ASSUMING A LINEAR RELATIONSHIP MODEL FOR EACH STAGE.

- $\text{Tau}_{\alpha w}$, $\text{Tau}_{\beta m}$

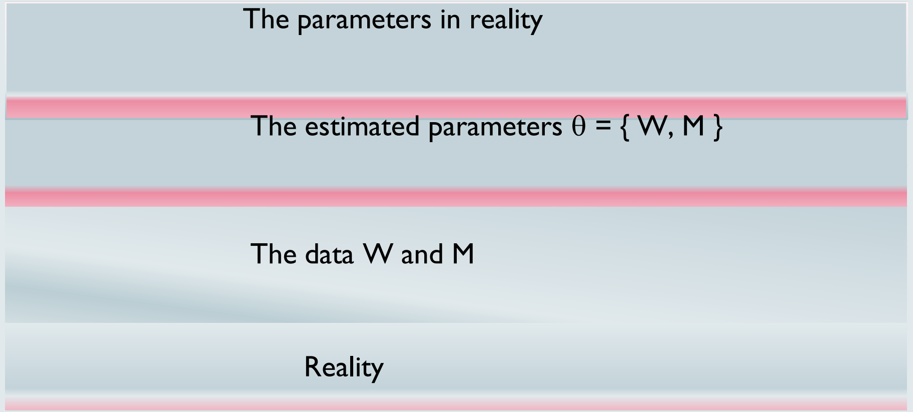
$\alpha_W, \beta_W, \alpha_M, \beta_M$

The risk of hiding her work

Data: n; w, m, x for i=1 to n

- Whether she says she works is M (0, 1). Whether she really works is W (0, 1).
- Wealth is X.

The known risk of working W based on a measure M



The parameters in reality

The estimated parameters θ = { W, M }

The data W and M

Reality

THIRD TYPE OF MODEL

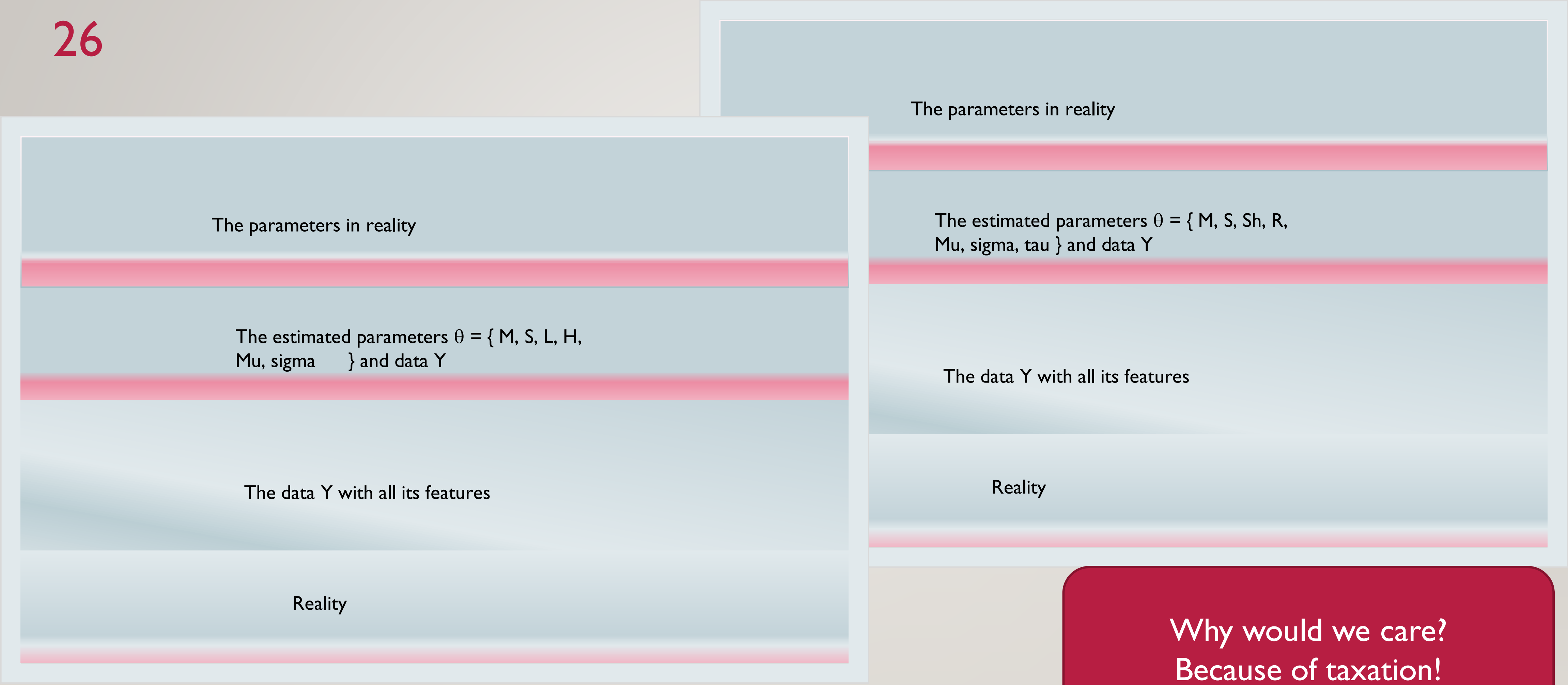
- In a different exercise, we have an outcome Y which is a continuous variable. As a random variable this has an expectation E(Y) and a variance V(Y) through observations i. Let us suppose the outcome is Y.
- We choose a prior function for the shape of the distribution of the parameter mu, again, such that E(pdf(mu)) = mean (Y) over many observations.
- In Bayesian Monte Carlo estimation we will simulate the distribution of E(pdf(Y)) and also E(Y) itself, assuming we are right about a posited functional from for the PDF of the hyperparameters and the parameters mu and var(Y)
- The prior and the maximum likelihood, together, are proportional to the **posterior distribution.** *The functional form we choose for the prior of hyperparameters will hardly affect the resulting distribution shape of PDF(Y), which has many moments. We can even allow PDF(Y) to have a shape which is a sum of a series of normal distributions, or a sum of individual binomial distributions, and the number of parameters can go up and up, along with hyperparameters like Tau(PDF(mu)). (sic. The dispersion of the estimate of the mean of the estimate of the mean)*
- The posterior distribution [best estimate], P({mu, tau, Mu(PDF(mu), Tau(PDF(mu))}| D), is not necessarily a symmetric function, nor a normal curve.
    - If you knew it was on a normal curve, that knowledge would be useful ‘prior’ information.
    - In such a case you can model it with a specific set of parameters, no skewness, no kurtosis, symmetric variation etc.

TRICKS

- Depending on what we know or surmise about the distribution of Y, we can set up different prior for this pair of hyperparameters:
- E(mu) and var(mu), where mu is the E(Y) from the PDF we posit for Y. (sic)
    - For example beta distribution if Y lies between 0 and 1 [rather than Tobit regression]
    - Or a lognormal distribution; Use a Poisson for PDF(Y) if it's a migrant count.
    - See Arek Wisniowski, paper on migrants into/out of UK+4 countries, J-RSS-A, 2016.
    - Bayesian theory shows that if Y is Poisson then the best hyperparameter distribution is Gamma
    - Alternatively you can always use uniform distributions.
- Another trick is that if we have 2 or 3 parameters, not just one, it's not just a 'Monte Carlo' Simulation- it's going to use a Gibbs Sampler [or Metropolis-Hastings] → MCMC

EXAMPLE 3: ESTIMATING THE DISTRIBUTION OF A CONTINUOUS VARIABLE (NOTES ATTACHED)

- Suppose we are really deeply interested in the distribution of earnings.
- We want to know more about the shape of the distribution.
- Here are two models of the distribution.



The parameters in reality

The estimated parameters θ = { M, S, L, H, Mu, sigma } and data Y

The data Y with all its features

Reality

The parameters in reality

The estimated parameters θ = { M, S, Sh, R, Mu, sigma, tau } and data Y

The data Y with all its features

Reality

Why would we care?
Because of taxation!

MCMC

- The Markov Chain is a series of estimates, similar to how a maximum likelihood solution is found by iteration.
- The Gibbs sampler first posits any estimate for mu, then estimates the posterior likelihood function for Tau from that, then moves to next step.
  - Now it uses the chosen Mu, either current or new, and posits a fresh new changed estimate of Tau, and generates the posterior likelihood function for Mu.
  - Each posterior likelihood function is conditional on the data and all the other parameters.
- As in bootstrapping, the Markov Chain of these estimates is getting better and better, closer to the true values. The var(theta) does not reduce to 0, but to the true s.d.(theta)

CONTINUATION OF LOGIC OF A CONTINUOUS SINGLE VARIABLE

- If you don't know it, then you may choose a flat prior, innocuous prior, or other specific prior for the distribution of Mu. Choosing this distribution may involve an implicit assumption about tau of mu. That is, we assert one distribution, and it implies a specific variance of the variable along that PDF.
- However, it is possible to find out that Y is normally distributed without assuming it is normally distributed. We do that by noting the hyperparametric nature of Mu of Y.
- Tau of Mu, and Mu itself, in particular can have any distribution, and so a 'credible interval' is wise choice.
- One credible interval will illustrate the distribution of Mu of Y which represents an estimate of E(Y), and does not represent the distribution of Y itself. But once we have the posterior distribution, we can also draw the entire distribution of Y itself, as long as we've built it up to enable it to have a variety of shapes.

CREDIBLE INTERVAL? OR HIGHEST PROBABILITY ZONE? OR HIGH DENSITY REGION

- These introductory authors call it the credible interval:

Kass, Robert E., and Adrian E. Raftery, 1995. **Bayes Factors**, *Journal of the American Statistical Association*, 90: 430. Review paper, 773-795.

- They explain the Bayes Information Criterion, a widely used measure of how good a model fit is, compared with an alternative model.
- They also explain the Akaike Information Criterion, which is very different in derivation, yet similar in its parameters.

AIC VS BIC

The Bayesian Information Criterion
BIC = -2(log maximised likelihood) + (log N)*(number of parameters)
Clearly, you try to minimise the BIC. But the N affects the BIC directly.
Also note, both AIC and BIC are penalising model complexity.
But the role of N is ambiguous.

- AIC= -2(log maximised likelihood) + 2(number of parameters)
- So you try to minimise AIC.
- See Lunn, Jackson, Best, Thomas, and Spiegelhalter, *The BUGS Book*,

pg 138, section 8.2, and pages 159-169.

- $D(\theta) = -2\log p(y|\theta)$ and

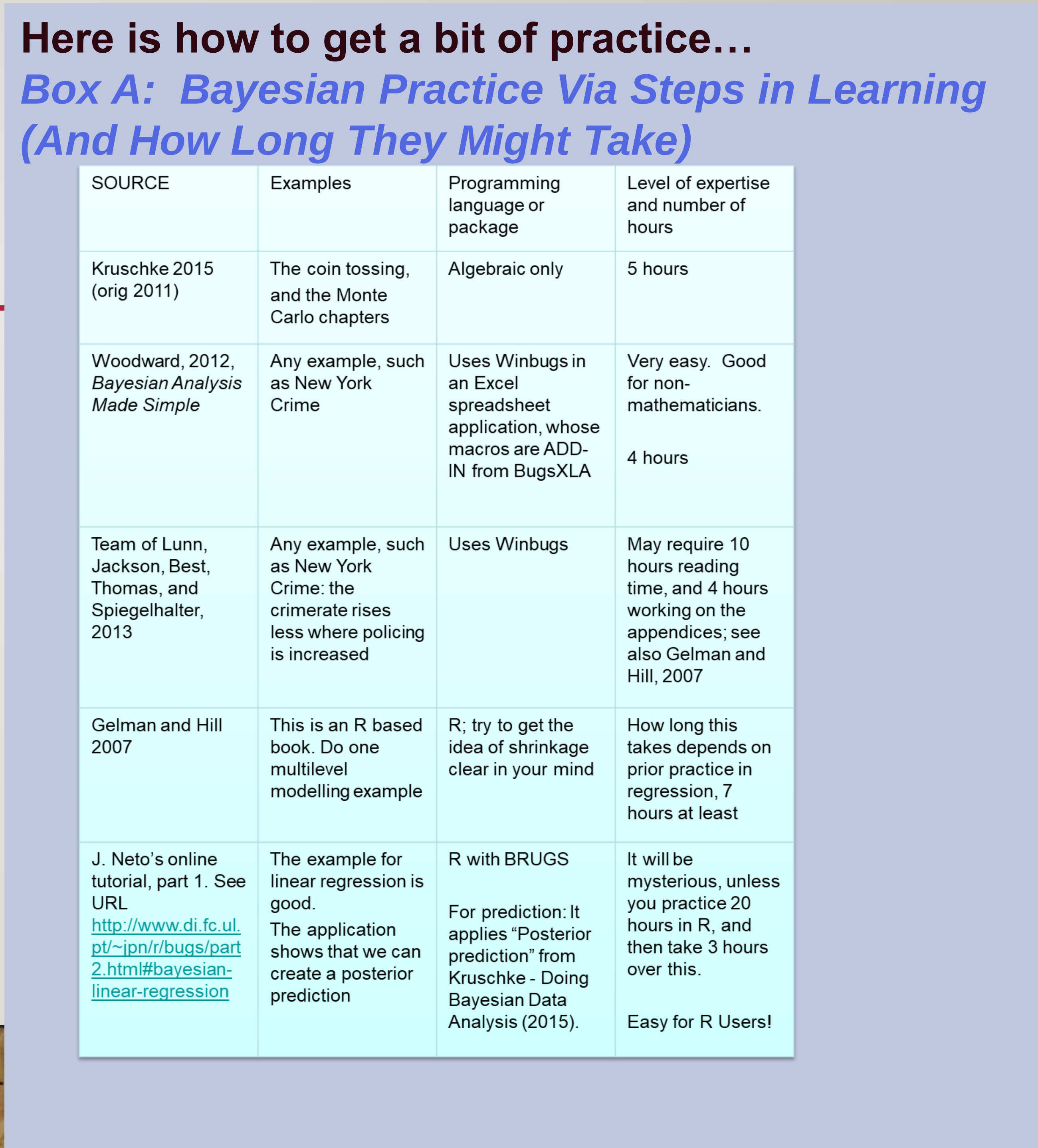
AIC = $-2 \log p(y|\hat{\theta}) + 2p = D(\hat{\theta}) + 2p$ Akaiki Information Criterion (pg 159)

- Here, the caret indicates maximum likelihood estimate, p is the dimension of θ, and is

the number of parameters in the whole model, and a lower AIC is favoured.

Why is log likelihood negative?
Because a small positive number has its log taken.

- Kruschke, John K. (2015), 2nd ed., *Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and STAN*, Amsterdam: Academic Press.



Here is how to get a bit of practice…

Box A: Bayesian Practice Via Steps in Learning (And How Long They Might Take)

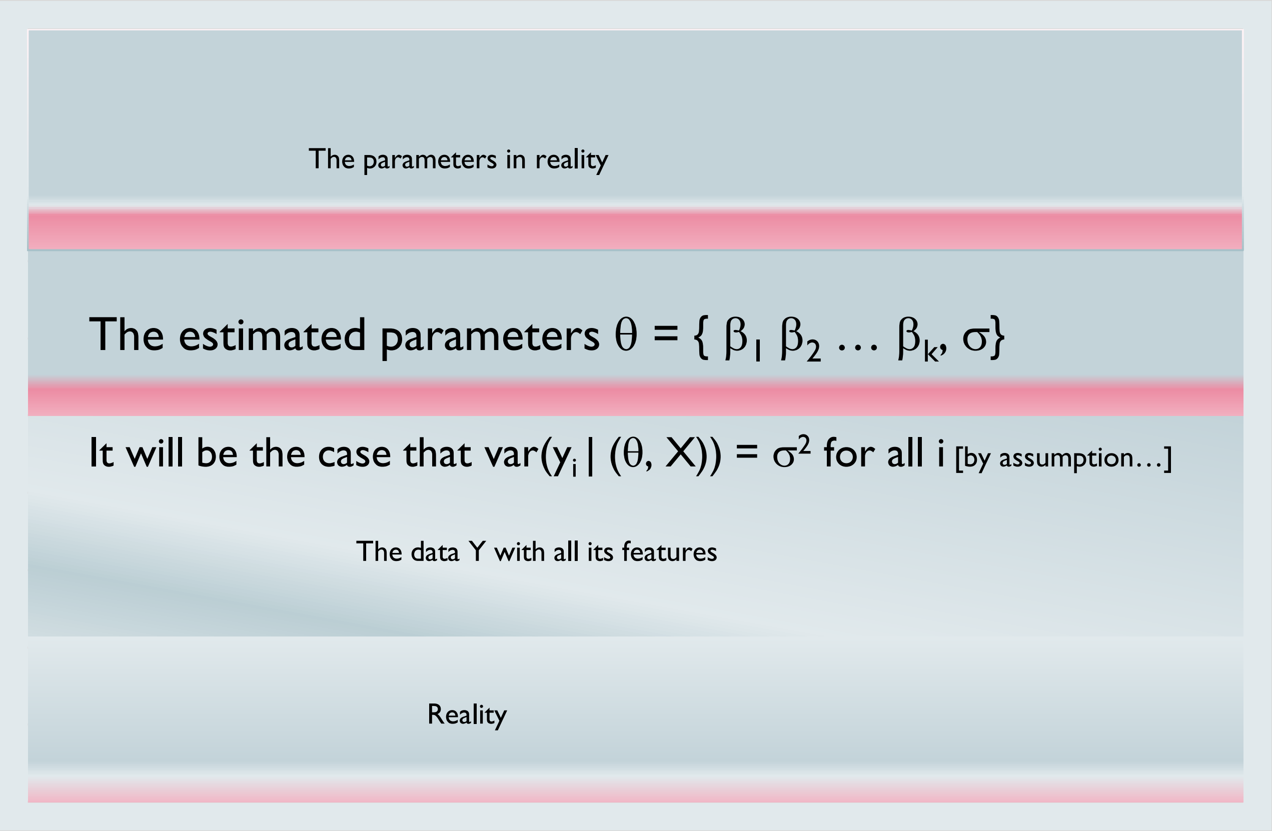
SOURCE	Examples	Programming language or package	Level of expertise and number of hours
Kruschke 2015 (orig 2011)	The coin tossing, and the Monte Carlo chapters	Algebraic only	5 hours
Woodward, 2012, *Bayesian Analysis Made Simple*	Any example, such as New York Crime	Uses Winbugs in an Excel spreadsheet application, whose macros are ADD-IN from BugsXLA	Very easy. Good for non-mathematicians. 4 hours
Team of Lunn, Jackson, Best, Thomas, and Spiegelhalter, 2013	Any example, such as New York Crime: the crimerate rises less where policing is increased	Uses Winbugs	May require 10 hours reading time, and 4 hours working on the appendices; see also Gelman and Hill, 2007
Gelman and Hill 2007	This is an R based book. Do one multilevel modelling example	R; try to get the idea of shrinkage clear in your mind	How long this takes depends on prior practice in regression, 7 hours at least
J. Neto's online tutorial, part 1. See URL http://www.di.fc.ul.pt/~jpn/r/bugs/part2.html#bayesian-linear-regression	The example for linear regression is good. The application shows that we can create a posterior prediction	R with BRUGS For prediction: It applies "Posterior prediction" from Kruschke - Doing Bayesian Data Analysis (2015).	It will be mysterious, unless you practice 20 hours in R, and then take 3 hours over this. Easy for R Users!

ADVANCED VERSION OF MODEL 3 - A REGRESSION IS POSITED DUE TO THE REAL CAUSAL MECHANISMS INVOLVED IN GENERATING OBSERVATIONS OF Y

- See Gelman and Hill, 2007, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press.
- Chapters 7 to 9. They say we can use simulation to estimate the linear regression model.

THE LINEAR MODEL AS SEEN IN HIERARCHICAL MODEL NOTATION

- The *ordinary linear model* assumes zero covariances of X_i with X_j but sigma, a level of variance, is not varying with X (the no-heteroskedasticity assumption). All cases are independently sampled.
- An *alternative linear model* would use logged wage as Y
- Many more alternatives exist



The parameters in reality

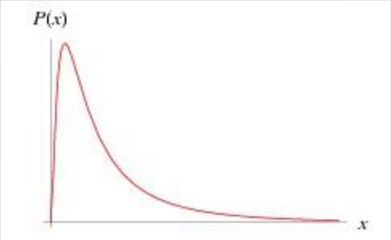
The estimated parameters $\theta = \{ \beta_1\ \beta_2 \ldots \beta_k, \sigma\}$

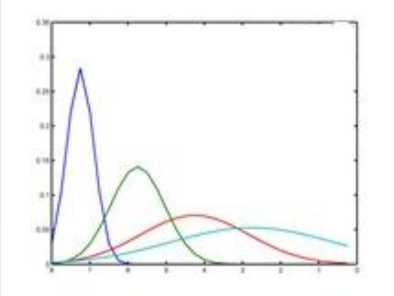
It will be the case that $\mathrm{var}(y_i \mid (\theta, X)) = \sigma^2$ for all i [by assumption…]

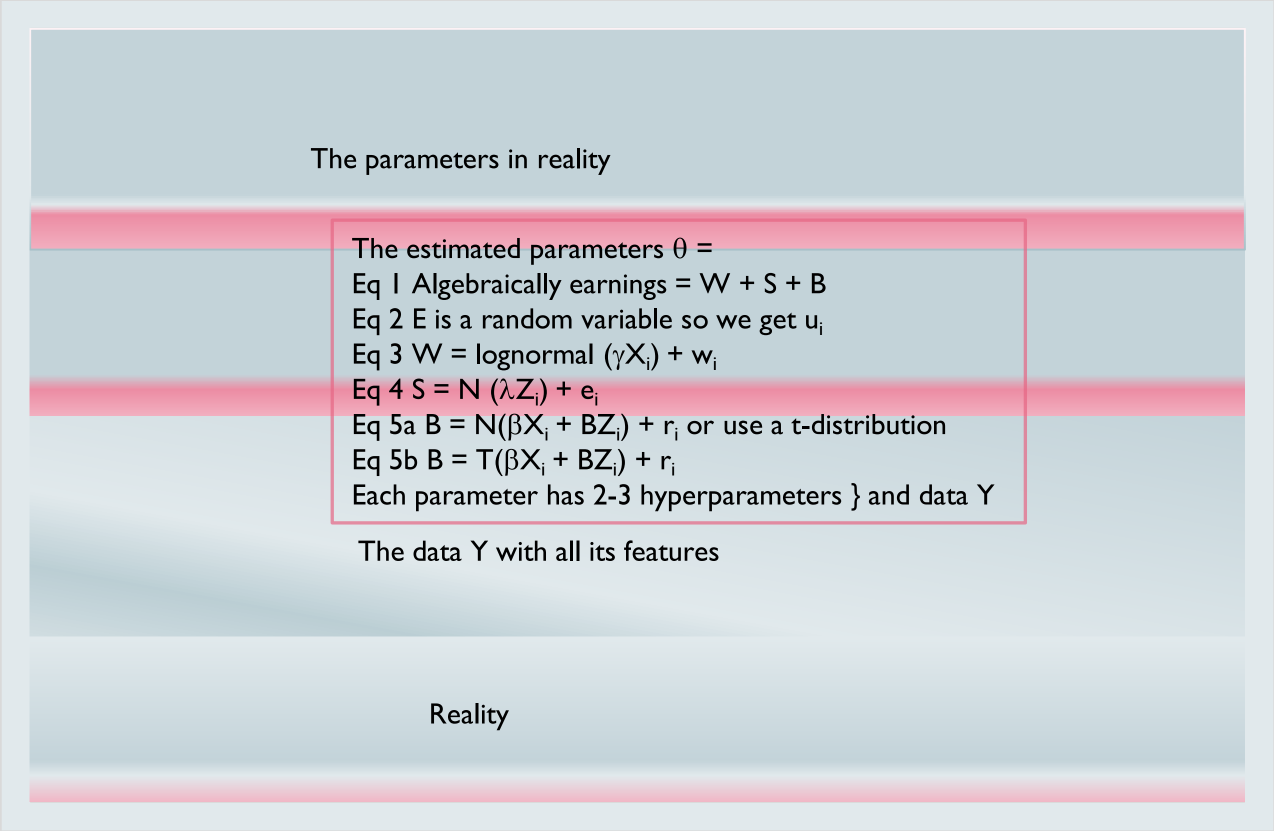
The data Y with all its features

Reality

EARNINGS MODEL







The parameters in reality

The estimated parameters θ =

Eq 1 Algebraically earnings = W + S + B

Eq 2 E is a random variable so we get u_i

Eq 3 $W = \text{lognormal}(\gamma X_i) + w_i$

Eq 4 $S = N(\lambda Z_i) + e_i$

Eq 5a $B = N(\beta X_i + BZ_i) + r_i$ or use a t-distribution

Eq 5b $B = T(\beta X_i + BZ_i) + r_i$

Each parameter has 2-3 hyperparameters } and data Y

The data Y with all its features

Reality

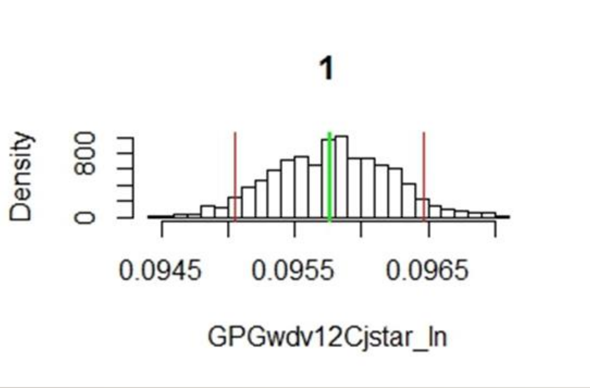
MODELLING POSSIBILITIES ARE LIMITLESS. INFORMATION FROM DELPHI METHOD CAN BE PUT INTO THE PROCESSING OF THE DATA.

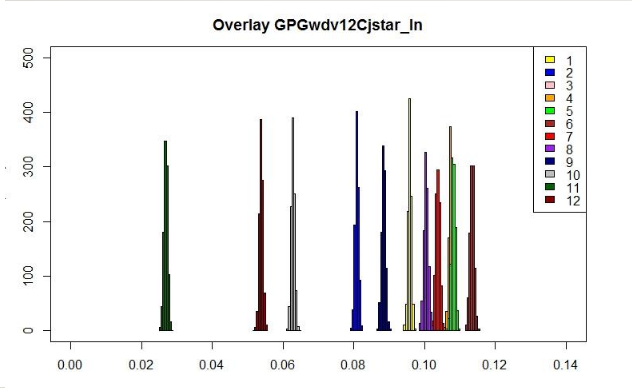
- Suppose there is variance within regions which is not equal, and variance between regions which helpfully disperses the overall variance of a continuous variable Y. Then regression should be multilevel. (Gelman and Hill, chapters 11-12, and ch. 14) A sociological analogy is that we could have social groups as Level 2.
- A Bayesian fit of the multilevel model begins with a factor for region. (or group)
- Then notice the equation (Gelman & Hill, page 279) illustrating that the rho of correlation of Y_{ij} between groups is estimated within the model parameter set Theta. The parameter set also includes the mean of all cases' intercepts and the mean of all ***regions***' slopes; the variance within the intercepts; and 2 covariances.

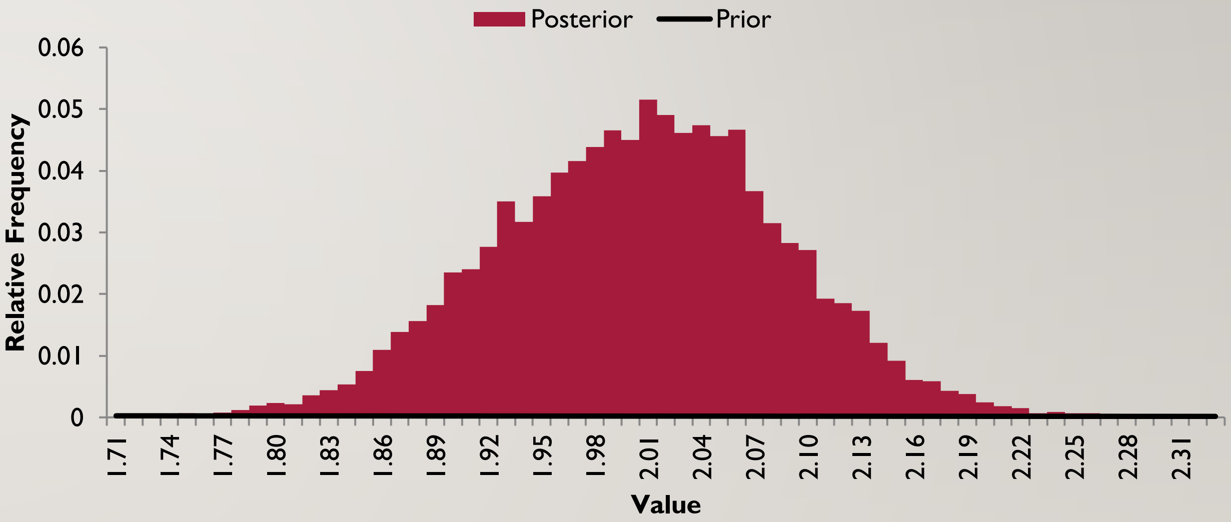
KEY COVARIANCES IN A MULTILEVEL MODEL

- The variance of Y has been parsed out into standard deviations of X and the error term
- In addition it breaks down into within and between,
- And we add the slope estimates for each region, beta
  - Which in turn have a variance(beta)
- And of course the intercepts of each region, alpha, also have:
  - Variance(alpha).
- The total number of parameters has risen considerably.

COMPARE THESE: A BOOTSTRAPPING ESTIMATE OF A SINGLE RANDOM VARIABLE, AT LEFT, VS. RESULTS OF ONE PARAMETER FROM A BAYESIAN ESTIMATION OF A LINEAR REGRESSION MODEL







Country Coefficient For Child Labour in Pakistan (Of Which, 2 Regions); Base=Nepal; Prior: Normal (mu=0 and sigma=17.9)

Figure 3: A Parameter with its Variation After MCMC Estimation

Methods of MCMC➔

We use Markov Chain Monte Carlo estimation to develop estimates of each of a set of parameters by looping through sets of possible candidate values, making a decision in an iterative manner, and revisiting the resulting model’s posterior probability, until the best possible set of parameter values has been reached

38 SUMMARY AND PRACTICE TASK

- **Summary: The Bayes Theorem says that**

$$\mathbf{P(\theta|D)} = \frac{\mathrm{P}(\mathbf{D}|\theta)\cdot\mathrm{P}(\theta)}{\mathrm{P}(\mathbf{D})}$$

$$\mathbf{P(\theta|D)} \propto \mathbf{P(D|\theta)}\cdot\mathrm{P}(\theta)$$

And $\mathbf{P(\theta|reality)} \leftarrow\rightarrow \mathbf{P}(\hat{\theta}|D)$

And $\mathbf{P}(\hat{\theta}|D) \propto$ **Likelihood · Prior**

Here we summarise. The Posterior is the basis for calculating the Y-hat distribution, but the posterior includes values for all parameters.

- **Task:**
- **Decide on a regression outcome, Ie a caused variable.**
- **Decide on 3 causal mechanisms which may affect Y.**
- **What distributional assumptions can you validly, vs. invalidly make?**

39 CONCLUSIONS

1. Using Bayesian models, the parameter estimates at the median may be equal to the estimates with Maximum Likelihood estimation. . I estimate these as the same up to 5 digits of accuracy in a wage-by-sex model.
2. The Bayes Factor also may be small. The Bayes Factor is the difference in the posterior probability of a hypothesis (e.g. B>0) under our model vs. a classical model. It can be used for non-nested models.
3. The standard deviation of an estimate may be narrow, if we use Bayesian methods, than in traditional methods!
4. But more importantly we can make a range of reasonable assumptions about it.
    1. I showed in a multilevel modelling context that the value of Y in a case may be correlated across regions, and this is not the same as assuming that all X variables are independent of each other (including the 'X variable' region). "Shrinkage" is the term for the relationship of regional means with the overall mean, compared with the old fashioned dummy variable method.
    2. You can manage your measurement error, such as Undercount, yourself!
5. The use of BIC makes more sense if we move to MCMC and hence Bayesian methods, but if you are using a big Structural Equation Model or MLM you will get a BIC anyway from an eclectic application of MCMC based on your classical statistical parameter set.
6. BIC exists if there is a likelihood estimate, which is maximised.
7. Bayesian Gibbs sampler methods give the same result but can they handle a much wider range of parameters.

ANSWER TO PRACTICAL TASK

- Y = f(W, M, L) = a+bW + cM + dL
- (a+bW + cM + dL) ~ N(μ_w, τ_w)
- M ~ N(μ_M, τ_M)
- L ~ … etc. L could be the rho representing the correlation of the outcome W and the M. We do not have to assume a zero correlation. It could have a range {-1, +1}.
- It may be on a Beta distribution.
- W in turn may depend on wealth S.

- Task:
- Decide on a regression outcome, Ie a caused variable.
- Decide on 3 causal mechanisms which may affect Y.
- What distributional assumptions can you validly, vs. invalidly make?
  - A) decide on key distributions of random variates.
  - B) decide on the covariance structure
  - C) use distribution diagrams Kruschke (