Bayesian Statistics in Accounting Research

EAA PhD Forum 2018 – Milan

Harm H. Schütt

LMU – Munich School of Managemet

May 30, 2018

Why Bayes?

"If I'm doing an experiment to save the world, I better use my prior." – Andrew Gelman

Disclaimer: I am no statistical theorist and I am still learning lots of things myself. This talk is supposed to be an applied perspective on which Bayesian tools are useful additions to our empirical tool belt.

Why Bayes?

"If I'm doing an experiment to save the world, I better use my prior." – Andrew Gelman

Disclaimer: I am no statistical theorist and I am still learning lots of things myself. This talk is supposed to be an applied perspective on which Bayesian tools are useful additions to our empirical tool belt.

Things we grapple with:

1. Noisy data, lots of correlated variables
   $\rightarrow$ Regularization
2. Sparse data, need to model latent constructs or heterogeneity
   $\rightarrow$ Hierarchical Models
3. Quantify uncertainty in estimates properly for decision making
   $\rightarrow$ Posterior Distribution

Agenda for Today

1. Quick overview of classical hypothesis testing
2. Brief introduction to Bayesian inference
3. Regularization (general and Bayesian adaptive versions)
4. Hierarchical Models / Model building
5. Summary and food for thought on applications in Accounting Research

Unfortunately no time for the ins and outs of fitting Bayesian models (But code for all examples/figures on Github at github.com/hschuett/EAA2018Bayes)

Classical Hypothesis Testing

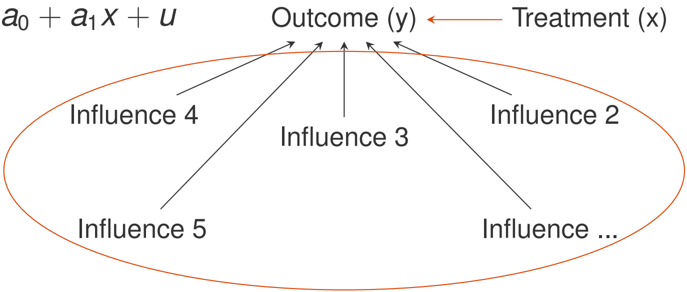
Key Problem of Inference

You want to learn something about an outcome from observations, but you cannot observe everything about the outcome

Key Problem of Inference

You want to learn something about an outcome from observations, but you cannot observe everything about the outcome

$$y = a_0 + a_1 x + u$$

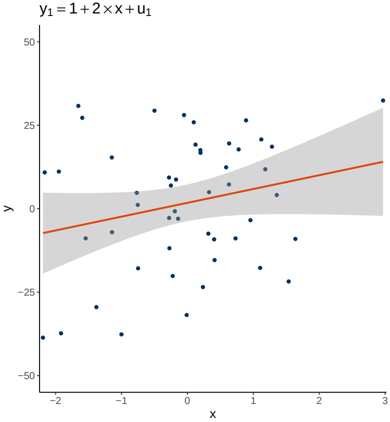


Unmeasured influences / “Noise” (u)

Nothing in the world is truly random – randomness is a vehicle to reason about unmeasured influences (e.g. a “random” coin toss)

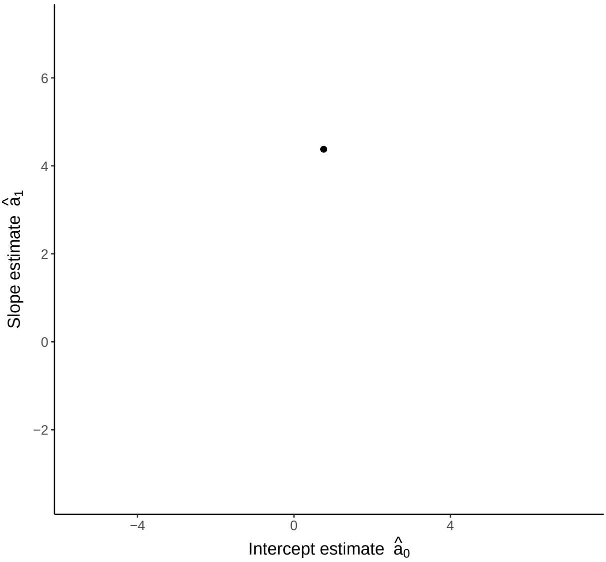
Unmeasured influences cause sampling variation

Two samples with the same $x \rightarrow y$ relation and the same x, produce different estimates because of unmeasured influences



So, how do you work with only one sample?

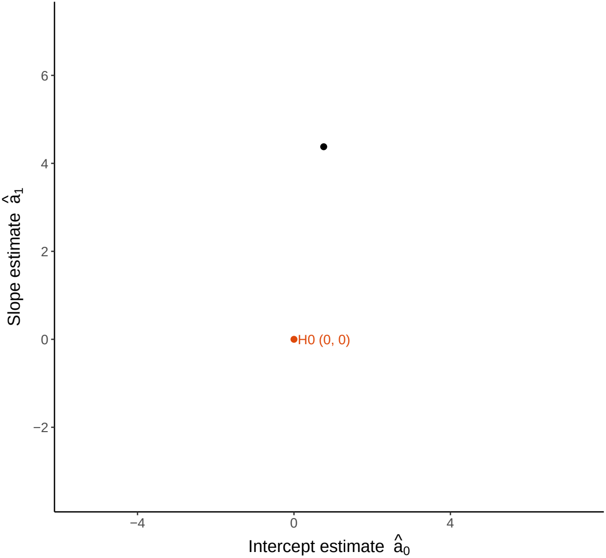
Inference: What are the true values a_0 and a_1?



Fitting $y = a_0 + a_1 x + u$

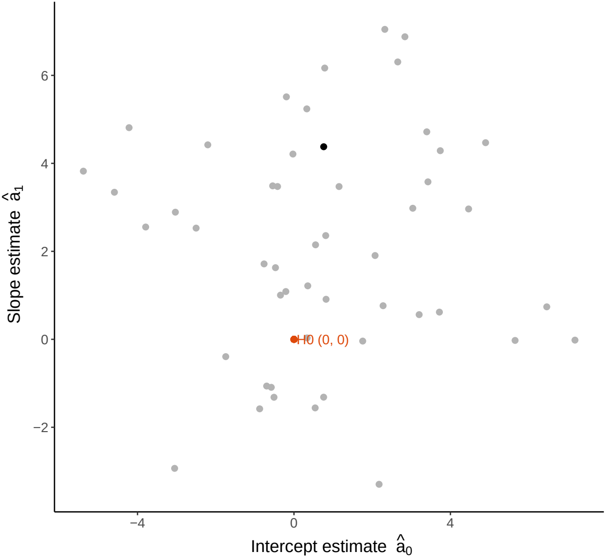
- Only one sample
- Estimates: $\hat{a}_0 = 0.8$, $\hat{a}_1 = 4.4$

Hypothesis Testing



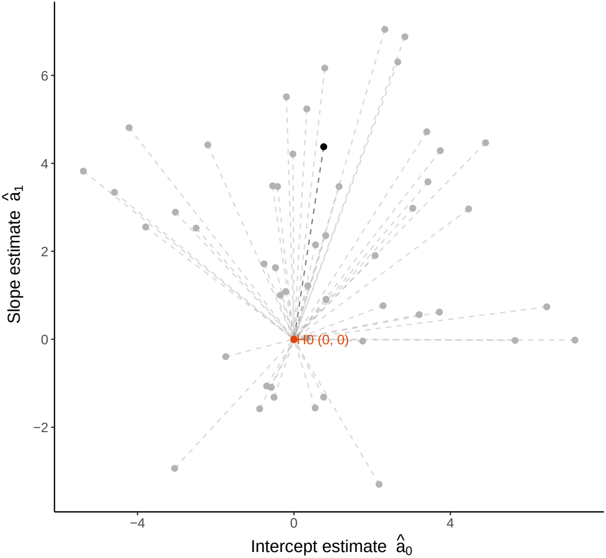
- Don't try to infer true value; Test a hypothesis *H*0
- How do we test *H*0?

Frequentist Uncertainty: Imaginary Resampling



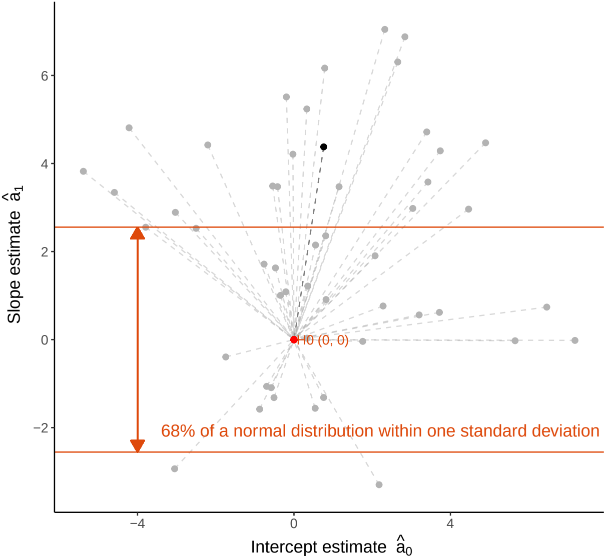
- Thought experiment. Suppose:
  1. H0 is true
  2. Regression setup correct
  3. Can draw more samples
- Constant x, variation comes from different u
- How often do we draw values like $(\hat{a}_0, \hat{a}_1)$?

Key Idea of an Hypothesis Test I



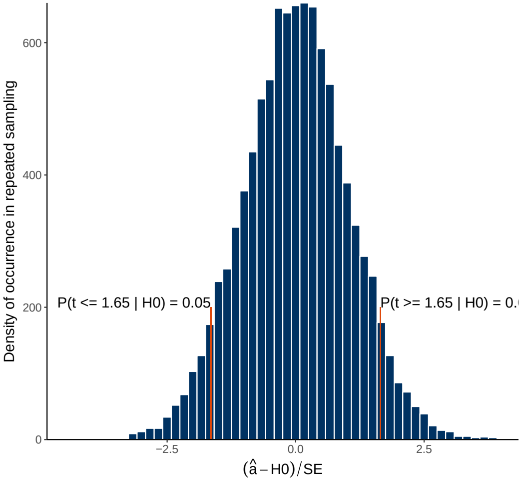
- The larger distance $\left|\hat{a}^1 - H0\right|$, the less likely $H0$
- But how unlikely?
- Can frame expected variation as a probability distribution (e.g., $\hat{a} \sim N(a, var[\hat{a}])$)
- All identification assumptions, error-term assumptions etc. are there to determine what distribution best reflects the expected variation (asymptotically)

Key Idea of an Hypothesis Test II



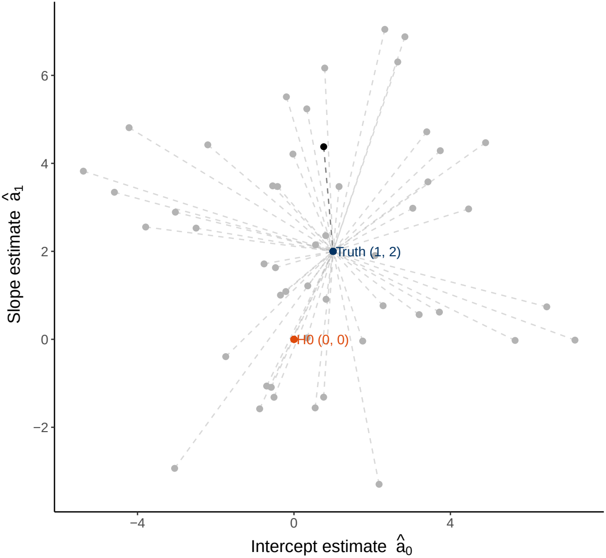
- Depending on your assumptions: $\hat{a} \sim N(a, var[\hat{a}])$
- Then
$$\frac{|\hat{a} - a|}{var[\hat{a}]} = N(0, 1)$$
- $var[\hat{a}]$ unobservable: you only have one sample and not many
- **Within sample estimate of** $var[\hat{a}]$**:** $SE(\hat{a})$

Key Idea of an Hypothesis Test III



- $\frac{|\hat{a}-a|}{var[\hat{a}]} = 1.65$
- Given expected Distribution, how often would 1.65 would occur if $H0$ true?
- Assumptions about u, which determine $SE(\hat{a})$ and form of test distribution

What are the True Values a_0 and a_1?



Solution: the true values

- $y = 1 + 2 * x + u$
- $\hat{a}_0^1 = 0.8$, $\hat{a}_1^1 = 4.4$
- Barely significantly different from zero at 10%
- True slope is 2, not 4.4
- Testing hypotheses; not inferring true values

So much noise that chance of estimates far away from true value is high

Summary

Great decision tool!

- (Only) two bits of data:
  - Estimate $\hat{a}$
  - $SE(\hat{a})$
- Fast and simple
- Effect of noise often unappreciated (i.e., "What does not kill it makes it stronger fallacy")
- Hard part is getting SEs correct and finding the test statistic distributions (i.e. asymptotic theory, etc.)

Bayesian Inference

Probability as a Measure of Plausibility

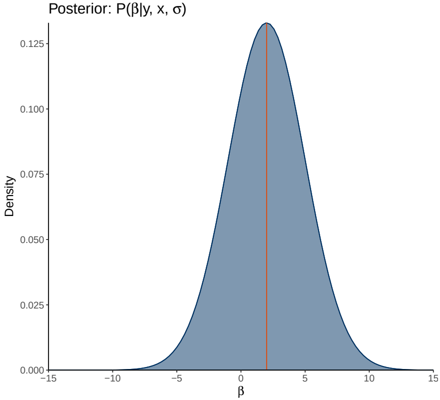
- Same setup: $y = X\beta + u$
- What if you want to know: what are the most plausible parameter values given the data?

Probability as a Measure of Plausibility

- Same setup: $y = X\beta + u$
- What if you want to know: what are the most plausible parameter values given the data?
- What you are looking for is the mode of the posterior distibution of the paramters given the data

$$P(\beta, \sigma | y, X)$$

But shape of $P(\beta, \sigma | y, X)$ also summarizes uncertainty!



OLS βs the Bayesian Way

Plausible parameter values given the data?

1. Formulate OLS model as a distribution:

$$y = X\beta + u \quad u \sim N(0, \sigma) \rightarrow y \sim N(X\beta, \sigma)$$

OLS βs the Bayesian Way

Plausible parameter values given the data?

1. Formulate OLS model as a distribution:

$$y = X\beta + u \quad u \sim N(0, \sigma) \rightarrow y \sim N(X\beta, \sigma)$$

2. The distribution gives a likelihood for the data y given unknown β and σ:

$$p(y|X, \beta, \sigma) = \phi(\frac{y - X\beta}{\sigma})$$

OLS βs the Bayesian Way

Plausible parameter values given the data?

1. Formulate OLS model as a distribution:

$$y = X\beta + u \quad u \sim N(0, \sigma) \rightarrow y \sim N(X\beta, \sigma)$$

2. The distribution gives a likelihood for the data y given unknown β and σ:

$$p(y|X, \beta, \sigma) = \phi(\frac{y - X\beta}{\sigma})$$

3. Add priors to ”turn the likelihood around”:

$$\underbrace{p\left(\beta, \sigma | y, X\right)}_{\text{Posterior}} = \frac{\overbrace{p\left(\beta, \sigma\right)}_{\text{Prior}} \overbrace{p\left(y | X, \beta, \sigma\right)}_{\text{Likelihood}}}{\underbrace{p\left(y | X\right)}_{\text{Data}}}$$

Difference Between Posterior and Likelihood

$$\underbrace{p(\beta,\sigma|y,X)}_{\text{Posterior}} = \frac{\underbrace{p(\beta,\sigma)}_{\text{Prior}}\underbrace{p(y|X,\beta,\sigma)}_{\text{Likelihood}}}{\underbrace{p(y|X)}_{\text{Data}}}$$

- Likelihood: $p(y|X,\beta,\sigma)$
  Probability of seeing the data y given parameters
- Posterior: $p(\beta,\sigma|y,X)$
  Probability of parameter being β given the data

If you want to know the most likely parameter given data, you are looking for the mode of the posterior, not the maximum likelihood.

Core Mechanism: Bayesian Updating

Example: we want to know the ability of a manager, measured as the probability θ of making good investments. Manager has a track record of $y = 8$ out of $n = 15$ successful investments

Core Mechanism: Bayesian Updating

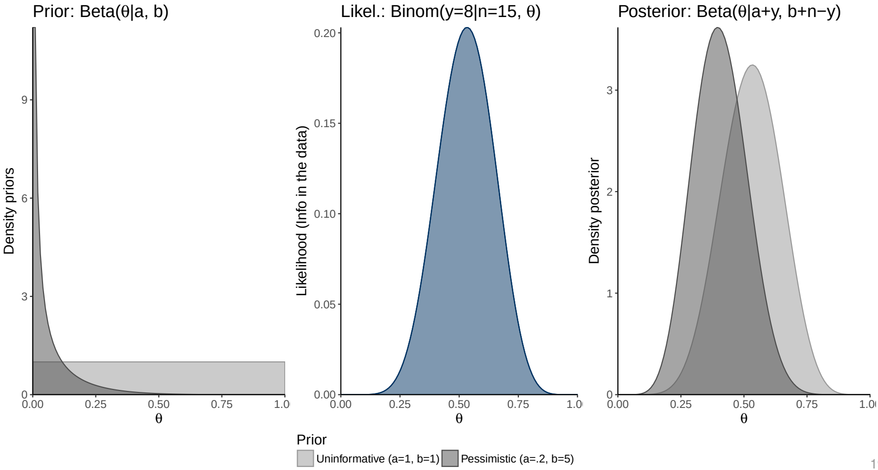
Example: we want to know the ability of a manager, measured as the probability θ of making good investments. Manager has a track record of $y = 8$ out of $n = 15$ successful investments

$$\underbrace{p(\theta)}_{\text{Prior}}\underbrace{p(y|n,\theta)}_{\text{Likelihood}} / \underbrace{p(y|n)}_{\text{Data}} = \underbrace{p(\theta|y,n)}_{\text{Posterior}}$$

Core Mechanism: Bayesian Updating

Example: we want to know the ability of a manager, measured as the probability θ of making good investments. Manager has a track record of $y = 8$ out of $n = 15$ successful investments

$$\underbrace{p(\theta)}_{\text{Prior}}\,\underbrace{p(y|n,\theta)}_{\text{Likelihood}}\,/\,\underbrace{p(y|n)}_{\text{Data}} = \underbrace{p(\theta|y,n)}_{\text{Posterior}}$$



Posterior and Prior

Posterior

- Outcome of interest
- Quantifies plausibility of different possible parameter values as a distribution

Posterior and Prior

Posterior

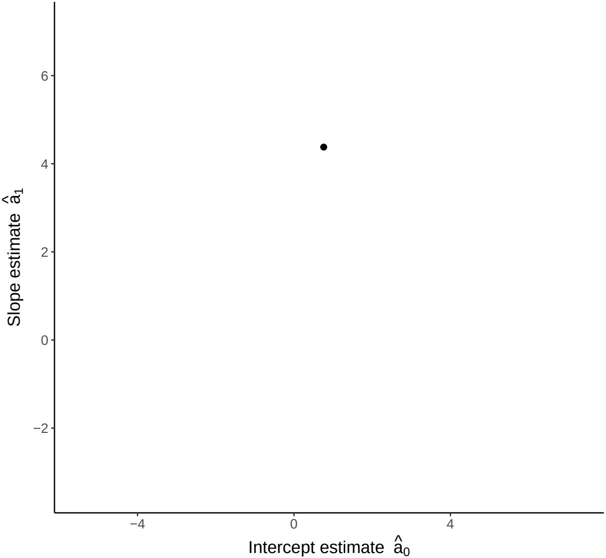
- Outcome of interest
- Quantifies plausibility of different possible parameter values as a distribution

Prior

- Input: Quantifies prior state of knowledge as a distribution
- Endless source of debate: how subjective, etc.
- Nowadays weakly informative priors quite common

Bayesian Regression

Back to our original OLS estimate



- $\hat{a}_0^1 = 0.8$ and $\hat{a}_1^1 = 4.4$
- How would a Bayesian version look like?

Bayesian Regression

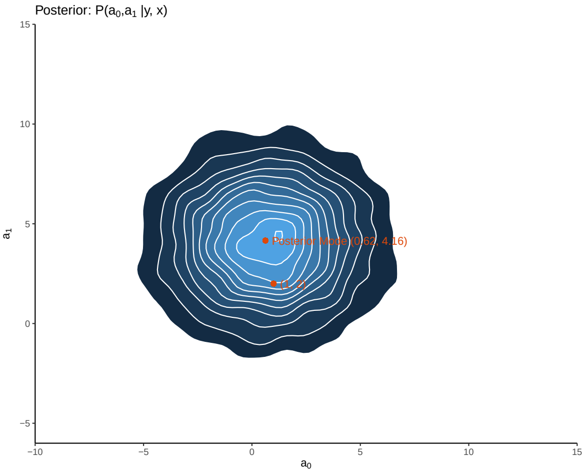
1. Specify a full model with distributions

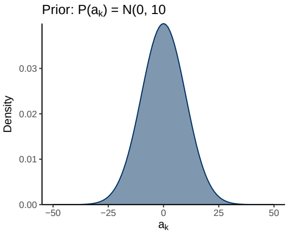
$$\begin{aligned} y &\sim N(\mu, \sigma) \\ \mu &= a_0 + a_1 * x \\ a_0 &\sim N(0, 10) \\ a_1 &\sim N(0, 10) \\ \sigma &\sim Cauchy(0, 20) \end{aligned}$$

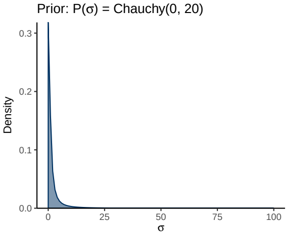
- Additional assumptions: The prior distributions
- Advantage: No more pain with inferences only holding asymptotically
- Advantage: You incorporate more information (which I argue is often a good thing)

Fitting $y = a_0 + a_1 * x + u$ with Very Weak Priors

Assume: We don't know much about what range the expected effects should be

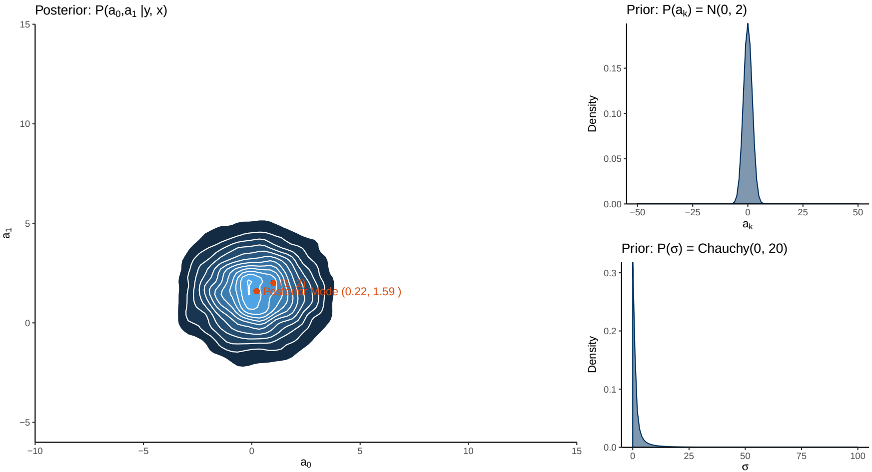






Fitting $y = a_0 + a_1 * x + u$ with Informative Priors

Assume: We are pretty confident that the effect, if present, shouldn't be bigger than 4



Why is Being Informative a Good Thing?

- Are you not distorting the data?

Why is Being Informative a Good Thing?

- Are you not distorting the data?
- Not if you believe in the prior. Does this introduce more subjectivity?

Why is Being Informative a Good Thing?

- Are you not distorting the data?
- Not if you believe in the prior. Does this introduce more subjectivity?
- From a practical perspective, I'd argue you often have not very debatable, weak priors that already help a lot combating noise/overfitting, etc.

Regularization

Regularization is Everywhere in Modern Statistics

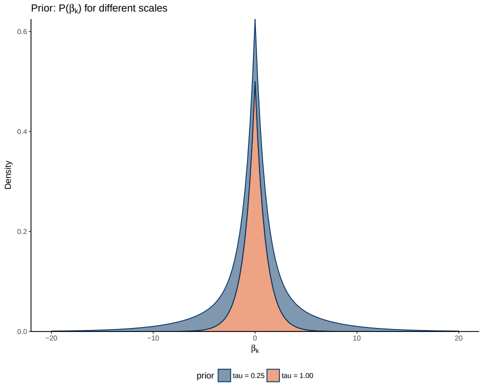
E.g., A Lasso Regression is nothing but a fast implementation of a Laplace prior

$$y \sim^{iid} N(X\beta, \sigma^2)$$
$$\beta \sim Laplace(0, \tau)$$

Regularization is Everywhere in Modern Statistics

E.g., A Lasso Regression is nothing but a fast implementation of a Laplace prior

$$y \sim^{iid} N(X\beta, \sigma^2)$$
$$\beta \sim Laplace(0, \tau)$$



then with $\lambda = \sigma^2/\tau$, the mode of $P(\beta|y, X, \sigma)$ minimizes

$$\underbrace{(y - X\beta)'(y - X\beta)}_{MeanSquaredError} + \underbrace{\lambda \sum_i |\beta_i|}_{LassoPenalty}$$

Regularizing Example: Has Speech Become More Polarized?

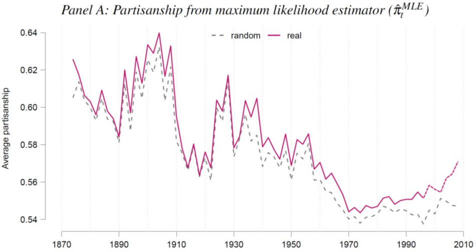
- Gentzkow et al. (2016): Measuring polarization in high-dimensional data: Method and application to congressional speech

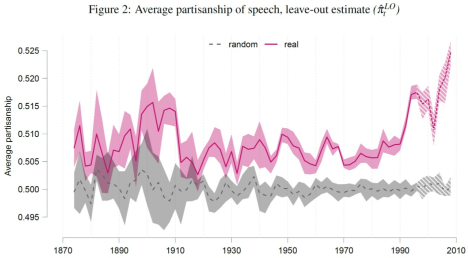
Regularizing Example: Has Speech Become More Polarized?

- Gentzkow et al. (2016): Measuring polarization in high-dimensional data: Method and application to congressional speech
- “Bias arises because the number of phrases a speaker could choose is large relative to the total amount of speech we observe, meaning many phrases are said mostly by one party or the other purely by chance” (p.3)

Regularizing Example: Has Speech Become More Polarized?

- Gentzkow et al. (2016): Measuring polarization in high-dimensional data: Method and application to congressional speech
- "Bias arises because the number of phrases a speaker could choose is large relative to the total amount of speech we observe, meaning many phrases are said mostly by one party or the other purely by chance" (p.3)
- Fig. 1 and 2:





Regularization in Modern Applications

- Many frequentist approaches like Lasso are fast, point estimate regularization implementations (e.g., Gentzkow et al., 2016)

Regularization in Modern Applications

- Many frequentist approaches like Lasso are fast, point estimate regularization implementations (e.g., Gentzkow et al., 2016)
- Essential to combating overfitting in Big-Data settings with a large number of correlated variables (Tibshirani, 2011, e.g.,)

Regularization in Modern Applications

- Many frequentist approaches like Lasso are fast, point estimate regularization implementations (e.g., Gentzkow et al., 2016)
- Essential to combating overfitting in Big-Data settings with a large number of correlated variables (Tibshirani, 2011, e.g.,)
- Similar use for variable selection. E.g., “We show that 15 out of 49 variables used in prior work robustly explain IPO returns.” (Butler et al., 2014, e.g.,)

Regularization in Modern Applications

- Many frequentist approaches like Lasso are fast, point estimate regularization implementations (e.g., Gentzkow et al., 2016)
- Essential to combating overfitting in Big-Data settings with a large number of correlated variables (Tibshirani, 2011, e.g.,)
- Similar use for variable selection. E.g., “We show that 15 out of 49 variables used in prior work robustly explain IPO returns.” (Butler et al., 2014, e.g.,)
- Bayesian approach the most flexible (Sometimes does not scale easily)

Automatic Regularization via Hierarchical Priors

Imagine you are interested in estimating some fixed effects

$$
\begin{aligned}
y_{i,t} &= a_i + u_{i,t} \\
a_i &\sim N(2, \sigma_a = 3) \\
u_{i,t} &\sim N(0, \sigma_u = 15)
\end{aligned}
$$

- Think of: CEO effects, firm effects, analyst effects
- Often only few observations per i (i.e., 5 years per firm)
- Use prior to discipline the a_i estimates

Regularization via Priors

Data Generating Process:

$$\begin{aligned} y_{i,t} &= a_i + u_{i,t} \\ a_i &\sim N(2, \sigma_a = 3) \\ u_{i,t} &\sim N(0, \sigma_u = 15) \end{aligned}$$

Bayesian model:

$$\begin{aligned} y_{i,t} &= a_i + u_{i,t} \\ a_i &\sim N(\mu_a, \sigma_a) \\ u_{i,t} &\sim N(0, \sigma) \\ \mu_a &\sim N(0, 100) \\ \sigma_a &\sim \mathit{HalfN}(0, 100) \\ \sigma_u &\sim \mathit{HalfN}(0, 100) \end{aligned}$$

Tiny bit of information helps already

Regularization via Priors

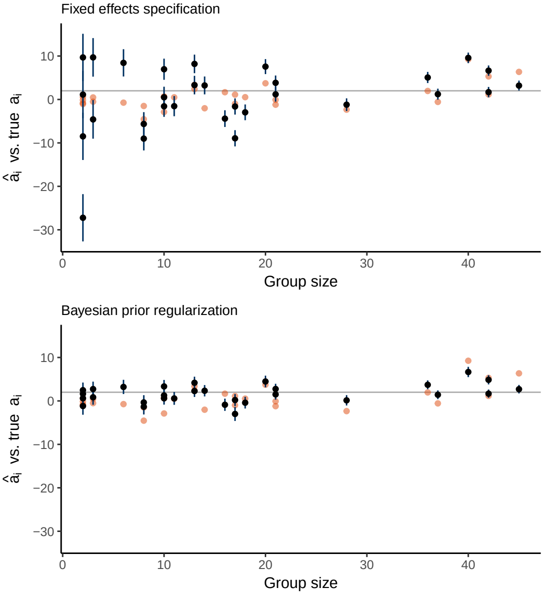
Data Generating Process:

$$
\begin{aligned}
y_{i,t} &= a_i + u_{i,t} \\
a_i &\sim N(2, \sigma_a = 3) \\
u_{i,t} &\sim N(0, \sigma_u = 15)
\end{aligned}
$$

Bayesian model:

$$
\begin{aligned}
y_{i,t} &= a_i + u_{i,t} \\
a_i &\sim N(\mu_a, \sigma_a) \\
u_{i,t} &\sim N(0, \sigma) \\
\mu_a &\sim N(0, 100) \\
\sigma_a &\sim HalfN(0, 100) \\
\sigma_u &\sim HalfN(0, 100)
\end{aligned}
$$

Tiny bit of information helps already



Hierarchical Models

Punchline so Far

- Bayesian inference rests on priors, which summarize prior information

Punchline so Far

- Bayesian inference rests on priors, which summarize prior information
- Regularization uses prior information to discipline estimates

Punchline so Far

- Bayesian inference rests on priors, which summarize prior information
- Regularization uses prior information to discipline estimates
- But many other important use cases for prior information:
    - Deal with sparse data, missing values
    - Combine multiple sources of data (chain them together via updating) to identify interesting constructs

Let's look at an example

The Best Part: Flexible Hierarchical Models

“Bayesian Estimation of Population-Level Trends in Measures of Health Status”
Finucane et al. (2014), largest-ever analysis of metabolic risk factors and the first global analysis of trends

The Best Part: Flexible Hierarchical Models

“Bayesian Estimation of Population-Level Trends in Measures of Health Status”
Finucane et al. (2014), largest-ever analysis of metabolic risk factors and the first global analysis of trends

- “Despite cardiovascular diseases being the leading causes of death worldwide (Lozano et al., 2013), our understanding of their trends is almost entirely based on specific cohorts and communities” (p.18)

The Best Part: Flexible Hierarchical Models

"Bayesian Estimation of Population-Level Trends in Measures of Health Status"
Finucane et al. (2014), largest-ever analysis of metabolic risk factors and the first global analysis of trends

- "Despite cardiovascular diseases being the leading causes of death worldwide (Lozano et al., 2013), our understanding of their trends is almost entirely based on specific cohorts and communities" (p.18)
- 199 prior studies from various countries, years, with varying quality

The Best Part: Flexible Hierarchical Models

"Bayesian Estimation of Population-Level Trends in Measures of Health Status"
Finucane et al. (2014), largest-ever analysis of metabolic risk factors and the first global analysis of trends

- "Despite cardiovascular diseases being the leading causes of death worldwide (Lozano et al., 2013), our understanding of their trends is almost entirely based on specific cohorts and communities" (p.18)
- 199 prior studies from various countries, years, with varying quality
- For roughly one-third of all countries, no data exist at all!

The Best Part: Flexible Hierarchical Models

“Bayesian Estimation of Population-Level Trends in Measures of Health Status”
Finucane et al. (2014), largest-ever analysis of metabolic risk factors and the first global analysis of trends

- “Despite cardiovascular diseases being the leading causes of death worldwide (Lozano et al., 2013), our understanding of their trends is almost entirely based on specific cohorts and communities” (p.18)
- 199 prior studies from various countries, years, with varying quality
- For roughly one-third of all countries, no data exist at all!

Bayesian model to combine data from all 199 studies, drawing strength from countries with data

A Hierarchical Model for Blood Pressure

$$y_{h,i} \sim N\left(\underbrace{a_{j[i]} + b_{j[i]} t_i + u_{j[i],t_i}}_{\text{country average + time trend}} + \overbrace{X_i' \beta + \gamma_i(z_h) + e_i}^{\text{age trends, study-level covariates}}, \frac{s_{h,i}^2}{n_{h,i} + \tau_i^2} \right)$$

$$a_j = a_{country} + a_{subregion} + a_{region} + a_{global}$$

$$b_j = b_{country} + b_{subregion} + b_{region} + b_{global}$$

$$a_c \sim N(0, \kappa_c), a_s \sim N(0, \kappa_s), a_r \sim N(0, \kappa_r), a_g \sim N(0, \kappa_g)$$

$$b_c \sim N(0, \eta_c), b_s \sim N(0, \eta_s), b_r \sim N(0, \eta_r), b_g \sim N(0, \eta_g)$$

y: blood pressure, i: study, j: country, h: gender

A Hierarchical Model for Blood Pressure

$$y_{h,i} \sim N\left(\underbrace{a_{j[i]} + b_{j[i]}t_i + u_{j[i],t_i}}_{\text{country average + time trend}} + \overbrace{X_i'\beta + \gamma_i(z_h) + e_i}^{\text{age trends, study-level covariates}}, \frac{s_{h,i}^2}{n_{h,i} + \tau_i^2}\right)$$

$$a_j = a_{country} + a_{subregion} + a_{region} + a_{global}$$
$$b_j = b_{country} + b_{subregion} + b_{region} + b_{global}$$
$$a_c \sim N(0,\kappa_c), a_s \sim N(0,\kappa_s), a_r \sim N(0,\kappa_r), a_g \sim N(0,\kappa_g)$$
$$b_c \sim N(0,\eta_c), b_s \sim N(0,\eta_s), b_r \sim N(0,\eta_r), b_g \sim N(0,\eta_g)$$

y: blood pressure, i: study, j: country, h: gender

- a_j, b_j structure assumes that countries close to each other (same subregion, etc.) are similar

A Hierarchical Model for Blood Pressure

$$y_{h,i} \sim N\left(\underbrace{a_{j[i]} + b_{j[i]} t_i + u_{j[i],t_i}}_{\text{country average + time trend}} + \overbrace{X_i' \beta + \gamma_i(z_h) + e_i}^{\text{age trends, study-level covariates}}, \frac{s_{h,i}^2}{n_{h,i} + \tau_i^2} \right)$$

$$a_j = a_{country} + a_{subregion} + a_{region} + a_{global}$$
$$b_j = b_{country} + b_{subregion} + b_{region} + b_{global}$$
$$a_c \sim N(0, \kappa_c), a_s \sim N(0, \kappa_s), a_r \sim N(0, \kappa_r), a_g \sim N(0, \kappa_g)$$
$$b_c \sim N(0, \eta_c), b_s \sim N(0, \eta_s), b_r \sim N(0, \eta_r), b_g \sim N(0, \eta_g)$$

y: blood pressure, i: study, j: country, h: gender

- a_j, b_j structure assumes that countries close to each other (same subregion, etc.) are similar
- Even if a country has no study, info about subregion AND info about distribution of country effects (e.g., $a_c \sim N(0, \kappa_c)$)!

Benefits of Hierarchical Structure

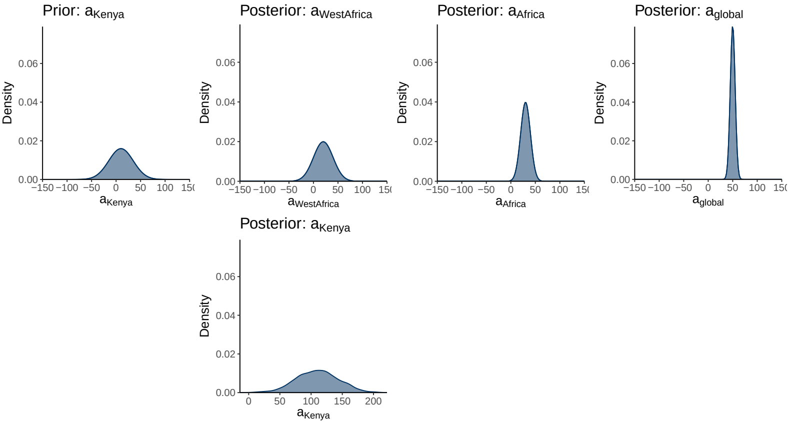
- Assume there is no study conducted about Kenya in a 30 year period? Can we say something about Kenya's blood pressure trends?

Benefits of Hierarchical Structure

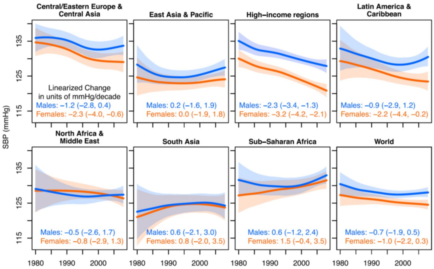
- Assume there is no study conducted about Kenya in a 30 year period? Can we say something about Kenya's blood pressure trends?
- Yes, the structure helps: $a_j = a_{country} + a_{subregion} + a_{region} + a_{global}$

Benefits of Hierarchical Structure

- Assume there is no study conducted about Kenya in a 30 year period? Can we say something about Kenya's blood pressure trends?
- Yes, the structure helps: $a_j = a_{country} + a_{subregion} + a_{region} + a_{global}$
- Chain evidence together:



Models with Impact



The results informed:

- WHO Global Status Report on noncommunicable diseases (NCDs; WHO, 2011)
- Targets for cardiovascular disease risk factors for the UN high-level meeting on NCDs
- US National Academy of Sciences Panel on International Health Differences in High Income Countries (Woolf and Aron, 2013)

Food for Thought
Use cases in Accounting

Firm-Level Heterogeneity

Let's take earnings "Persistence" as an important Accounting firm-level concept.

$$RoA_{i,t+1} = \beta_i RoA_{i,t} + u_{i,t}$$

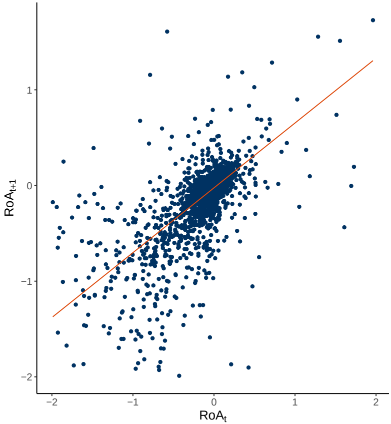
How do we usually estimate this?

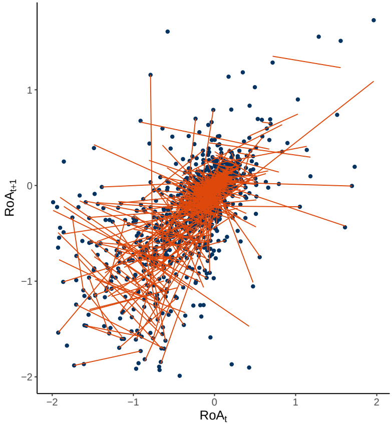
Firm-Level Heterogeneity

Let's take earnings "Persistence" as an important Accounting firm-level concept.

$$RoA_{i,t+1} = \beta_i RoA_{i,t} + u_{i,t}$$

How do we usually estimate this?





Firm-Level Heterogeneity

Can we do better? Let's frame this in a Bayesian model and use weak adaptive priors to regularize

$$
\begin{aligned}
RoA_{i,t+1} &\sim N(\beta_i RoA_{i,t}, \sigma_u) \\
\beta_i &\sim N(\mu_{firm}, \sigma_{firm}) \\
\mu_{firm} &\sim N(0, 10) \\
\sigma_{firm} &\sim HalfN(0, 10) \\
\sigma_u &\sim HalfN(0, 10)
\end{aligned}
$$

How much shrinkage Do we get?

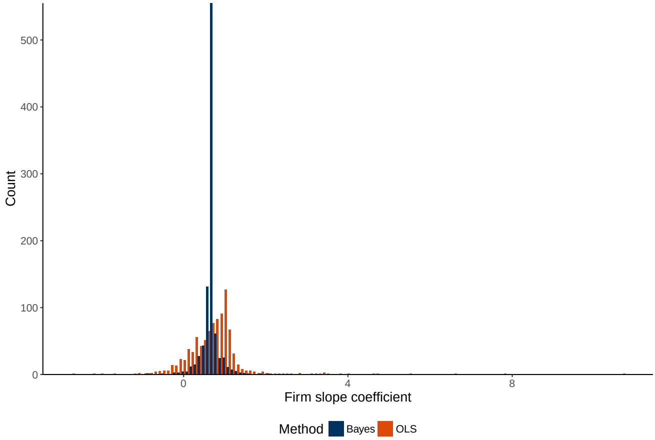
Firm-Level Heterogeneity

Can we do better? Let's frame this in a Bayesian model and use weak adaptive priors to regularize

$$\begin{aligned} RoA_{i,t+1} &\sim N(\beta_i RoA_{i,t}, \sigma_u) \\ \beta_i &\sim N(\mu_{firm}, \sigma_{firm}) \\ \mu_{firm} &\sim N(0, 10) \\ \sigma_{firm} &\sim HalfN(0, 10) \\ \sigma_u &\sim HalfN(0, 10) \end{aligned}$$

How much shrinkage Do we get?

Method	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
OLS	-48.0034	0.3493	0.7480	0.6252	0.9985	10.6649
Bayes	-0.9404	0.6479	0.6722	0.6703	0.7059	2.1198



Posterior Means:

- $\mu_{firm} = 0.67$
- $\sigma_{firm} = 0.04$
- $\sigma_u = 0.19$

Settings in Accounting Research

- Noisy settings, sparse data
  e.g., public/private firm comparisons, management effects, text data

Settings in Accounting Research

- Noisy settings, sparse data
  e.g., public/private firm comparisons, management effects, text data
- Large number of latent, hard to measure constructs
  e.g., accrual quality, audit quality, investment efficiency, financial constraints, etc. (Example: inferring news outlet audience characteristics (Schütt, 2018))

Settings in Accounting Research

- Noisy settings, sparse data
  e.g., public/private firm comparisons, management effects, text data
- Large number of latent, hard to measure constructs
  e.g., accrual quality, audit quality, investment efficiency, financial constraints, etc. (Example: inferring news outlet audience characteristics (Schütt, 2018))
- Modelling Heterogeneity
  So far, we are mainly concerned with average effects

Settings in Accounting Research

- Noisy settings, sparse data
  e.g., public/private firm comparisons, management effects, text data
- Large number of latent, hard to measure constructs
  e.g., accrual quality, audit quality, investment efficiency, financial constraints, etc. (Example: inferring news outlet audience characteristics (Schütt, 2018))
- Modelling Heterogeneity
  So far, we are mainly concerned with average effects
- Properly accounting for uncertainty
  Two stage procedures need to explicitly adjust standard errors to propagate uncertainty. A Bayesian model does this automatically

Settings in Accounting Research

- Noisy settings, sparse data
  e.g., public/private firm comparisons, management effects, text data
- Large number of latent, hard to measure constructs
  e.g., accrual quality, audit quality, investment efficiency, financial constraints, etc. (Example: inferring news outlet audience characteristics (Schütt, 2018))
- Modelling Heterogeneity
  So far, we are mainly concerned with average effects
- Properly accounting for uncertainty
  Two stage procedures need to explicitly adjust standard errors to propagate uncertainty. A Bayesian model does this automatically
- Combing multiple data sources
  e.g., survey and archival data (Work in progress)

Summary of Advantages

- In Frequentist Statistics, only measurements can have distributions, in Bayesian Statistics data and parameter have distributions
- If data is missing, it can be viewed as a parameter to be estimated

Summary of Advantages

- In Frequentist Statistics, only measurements can have distributions, in Bayesian Statistics data and parameter have distributions
- If data is missing, it can be viewed as a parameter to be estimated
- Clever setup of hierarchical models allows us to exploit commonalities
- Uncertainty at every level of the model is accounted for. No need for adjusting standard errors, etc

Summary of Advantages

- In Frequentist Statistics, only measurements can have distributions, in Bayesian Statistics data and parameter have distributions
- If data is missing, it can be viewed as a parameter to be estimated
- Clever setup of hierarchical models allows us to exploit commonalities
- Uncertainty at every level of the model is accounted for. No need for adjusting standard errors, etc
- Easy to model many forms of treatment heterogeneity (You can have more parameters than observations)

Summary of Advantages

- In Frequentist Statistics, only measurements can have distributions, in Bayesian Statistics data and parameter have distributions
- If data is missing, it can be viewed as a parameter to be estimated
- Clever setup of hierarchical models allows us to exploit commonalities
- Uncertainty at every level of the model is accounted for. No need for adjusting standard errors, etc
- Easy to model many forms of treatment heterogeneity (You can have more parameters than observations)
- Shape of posterior distribution ideal measure of uncertainty

Are There Disadvantages?

Yes. Computational complexity

- Most posterior distributions cannot be derived analytically. Estimated by brute force Markov Chain Monte Carlo Simulation.
- Models with a lot of data and/or large number of parameters can take a long time to fit (i.e., days)
- If something is wrong with the model, it is harder to figure out what went wrong
- But: Very active area of research in speeding things up (CUDA MCMC, Variational Inference, TensorFlow Probability, etc.)

Learning Bayesian Statistics

In case you got interested

Take a course!

Textbooks:

- Gelman et al. (2013)

Software:

- Stan. Biggest momentum, easiest to use (Pystan, Rstan), very active forum and lots of case studies (examples used in this presentation are coded with Rstan)
- JAGS. Older but well optimized and established sampler
- Nimble. For the real experts, who want to write their own samplers
- Stata modules available. I have no experience with them though
- TensorFlow Probability. For large scale machine learning

Thank you for your attention

References I

BUTLER, A. W., M. O. KEEFE, AND R. KIESCHNICK (2014): "Robust determinants of IPO underpricing and their implications for IPO research," *Journal of Corporate Finance*, 27, 367–383.

FINUCANE, M. M., C. J. PACIOREK, G. DANAEI, AND M. EZZATI (2014): "Bayesian estimation of population-level trends in measures of health status," *Statistical Science*, 18–25.

GELMAN, A., J. B. CARLIN, H. S. STERN, D. B. DUNSON, A. VEHTARI, AND D. B. RUBIN (2013): *Bayesian data analysis*, Chapman & Hall/CRC, 3 ed.

GENTZKOW, M., J. M. SHAPIRO, AND M. TADDY (2016): "Measuring polarization in high-dimensional data: Method and application to congressional speech," Tech. rep., National Bureau of Economic Research.

SCHÜTT, H. H. (2018): "Competition in Financial News Markets and Trading Activity," Available at SSRN:, available at SSRN:.

References II

TIBSHIRANI, R. (2011): "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 273–282.