PLSC 502 – Fall 2024
Miscellanea

December 9, 2024

Missing Data

Missing Data

Why missing?

- The observation doesn't exist
- The data don't exist for that observation
- The data exist, but are *impossible* to measure
- The data exist, but were not measured

Key: **Understanding the "missingness mechanism."**

A Framework for Missing Data

Notation:

$$\underset{N\times k}{\mathbf{X}} \cup \{W, Z\}$$

W_i have some missing values,

Z_i are "complete"

Then think of a matrix $\underset{N\times k}{\mathbf{R}}$ with:

$$R_{ik} = \begin{cases} 1 & \text{if } W_{ik} \text{ is missing,} \\ 0 & \text{otherwise.} \end{cases}$$

Example: **X** and **R**

So for:

$$\mathbf{X} = \begin{bmatrix} X_1 & X_2 & X_3 & \cdots & X_k \\ 17.7 & 220 & \texttt{NA} & \cdots & 1 \\ 14.9 & \texttt{NA} & 1982 & \cdots & 1 \\ 21.1 & 160 & 1959 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 11.9 & \texttt{NA} & 2001 & \cdots & \texttt{NA} \end{bmatrix}$$

we have:

$$\mathbf{R} = \begin{bmatrix} R_1 & R_2 & R_3 & \cdots & R_k \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 1 \end{bmatrix}$$

Framework (continued)

Rubin's flavors of missingness:

- Missing completely at random ("MCAR") (a/k/a "ignorably" missing):

$$\mathbf{R} \perp \{Z, W\}$$

- Missing at random ("MAR") ("conditionally ignorable" missingness):

$$\mathbf{R} \perp W | Z$$

- Anything else is "informatively" (or "non-ignorably") missing (sometimes called "Not missing at random" / "NMAR")

Rubin's Flavors Remix

Suppose we have two variables, an outcome Y and a covariate / predictor X. Define $R_{(Y)}$ as the vector of missing data indicators for Y (analogously to above).

Then:

$$
\begin{tikzcd}
X \arrow[rr] & & Y \\
R_{(Y)} & &
\end{tikzcd}
$$

Missing Completely At Random (MCAR)

Rubin Remixed (continued)

$$\begin{tikzcd} X \arrow[rr] \arrow[d] & & Y \\ R_{(Y)} & & \end{tikzcd}$$

Missing At Random (MAR)

Rubin Remixed (continued)

$$
\begin{tikzcd}
X \arrow[r] & Y \arrow[dl] \\
R_{(Y)} &
\end{tikzcd}
$$

Not Missing At Random (NMAR)

Missingness Due To Confounding

$$\begin{tikzcd} X \arrow[r] & Y \\ R_{(Y)} & (U) \arrow[l] \arrow[u] \arrow[ul, dashed] \end{tikzcd}$$

(Also) Not Missing At Random (NMAR)

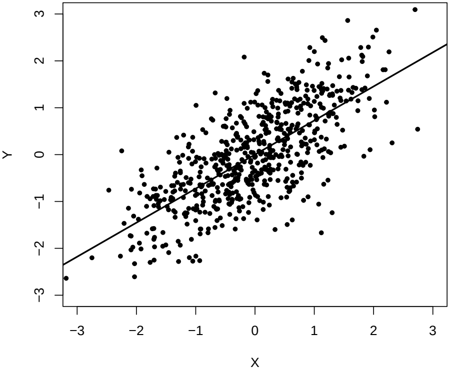
Missing Data Types, Part III

Suppose you have a survey of a simple random sample of $N = 500$ Penn State students, where questions include gender, age, and frequency of binge drinking.

Situation	Result
Your laptop randomly deletes 100 responses.	MCAR
Students who are underage are more likely not to respond.	MAR
Male students are more likely to binge drink *and* less likely to respond as a result.	NMAR
Students with fake IDs are less likely to respond *and* more likely to binge drink.	NMAR

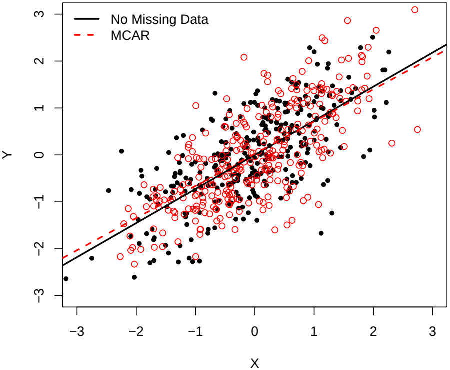
Simulated Illustration

```
> set.seed(7222009)
> data<-as.data.frame(rmvnorm(600,mean=c(0,0),
              sigma=matrix(c(1,0.707,0.707,1),nrow=2)))
> colnames(data)<-c("X","Y")
```



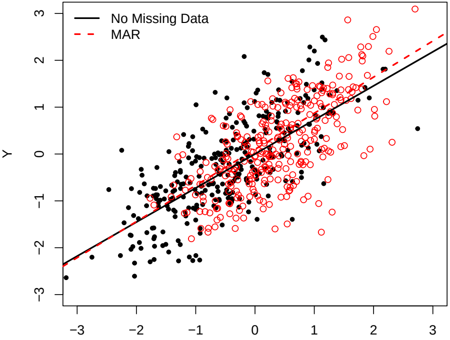
Example: MCAR

```
> # Flag 300 observations randomly for deletion:
>
> MCARs<-sample(nrow(data),300,replace=FALSE)
> data$MCAR<-ifelse(rownames(data) %in% MCARs,1,0)
>
> # Is the missingness related to X or Y?
>
> t.test(Y~MCAR,data=data)$statistic
     t
0.2268
> t.test(X~MCAR,data=data)$statistic
      t
-0.5848
```



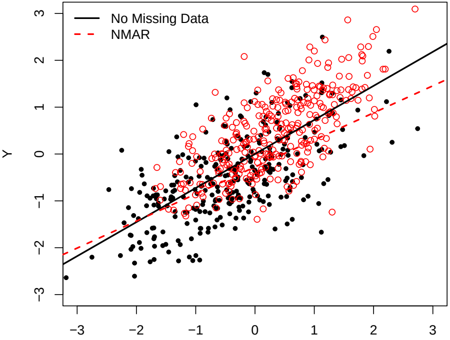
Example: MAR

```
> # Flag 300 observations for deletion based on their
> # values on X:
>
> MARs<-sample(nrow(data),300,replace=FALSE,prob=pnorm(data$X))
> data$MAR<-ifelse(rownames(data) %in% MARs,1,0)
>
> # Is the missingness related to X or Y?
>
> t.test(Y~MAR,data=data)$statistic
    t
-5.43
> t.test(X~MAR,data=data)$statistic
     t
-11.11
```



Example: NMAR

```
> # Flag 300 observations for deletion based on their
> # values on Y:
>
> NMARs<-sample(nrow(data),300,replace=FALSE,prob=pnorm(data$Y))
> data$NMAR<-ifelse(rownames(data) %in% NMARs,1,0)
>
> # Is the missingness related to X or Y?
>
> t.test(Y~NMAR,data=data)$statistic
     t
-13.21
> t.test(X~NMAR,data=data)$statistic
     t
-9.705
```



NMAR: Unmeasured Confounder

```
> # Flag 300 observations for deletion based on their values on an "unmeasured" variable U
> # that is related to X and Y:
>
> data$U <- (2*data$Y)-(2*data$X)
> NMAR2<-sample(nrow(data),300,replace=FALSE,prob=pnorm(data$U))
> data$NMAR2<-ifelse(rownames(data) %in% NMAR2,1,0)
>
> t.test(Y~NMAR2,data=data)$statistic
     t
-6.788
> t.test(X~NMAR2,data=data)$statistic
    t
2.074
```

Missing Data: What To Do?

- Listwise deletion / "complete cases analysis"
- Pairwise deletion / "available case analysis"
- Interpolation / replacement values / other static approaches
- *Imputation*-based approaches
  - Essentially, filling in missing values with "likely" values based on covariate patterns in the (observed) data
  - Usually done repeatedly, and average over results (hence, "multiple imputation")

Bayesian Statistics

"Frequentist" Statistics

Frequentist (sometimes, "objectivist") paradigm:

- Probability = <u>Long-run relative frequency</u>
- $\Pr(X)$ is a <u>fixed</u> but <u>unknown</u> quantity
- $\rightarrow$ the "single event problem"
  - Implies *repeatable* events
  - "The probability it will snow in State College, PA on December 9, 2024" is incoherent

Bayesian Probability

Bayesian (or, sometimes, "subjectivist"):

- Quantity of interest (θ)
- Data (X)
- sampling density [$\Pr(X|\theta)$]
- We want to know $\Pr(\theta|X)$
- Likelihood $L(\theta|X) \propto \Pr(X|\theta)$

Kolmogorov's Axioms + Conditional Probability

Kolmogorov requires that:

1. $\Pr(X) > 0$ (probabilities are non-negative)
2. $\Pr(X = a) \cup \Pr(X = b) \cup \Pr(X = c) ... = 1$ (the union of the probabilities of all possible outcomes equals one)
3. $\Pr(X = a) + \Pr(X = b) = \Pr(X = a) \text{ or } \Pr(X = b)$ (events are mutually exclusive)

If these are true, then for two events A and B:

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

Bayes' Rule

The rule for conditional probability also implies that:

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

So:

$$\Pr(A \cap B) = \Pr(B|A)\Pr(A)$$

Substituting, we get *Bayes' Rule*:

$$\overbrace{\Pr(A|B)}^{\textit{Posterior}} = \frac{\overbrace{\Pr(B|A)}^{\textit{Likelihood}}\overbrace{\Pr(A)}^{\textit{Prior}}}{\underbrace{\Pr(B)}_{\textit{Data/Evidence}}}$$

Example: Base Rates Blues

Suppose that:

- 1 in 100 undergraduates cheats on a paper assignment, and
- Your plagiarism detection software detects 100 percent of cheating, but
- 1 out of 100 times gives a "false positive" (i.e., indicates that the student cheated when they did not).

If the software flags a paper for plagiarism (P), what is the probability that the student cheated (C)?

We have:

$$
\begin{aligned}
\Pr(P|C) &= 1 \\
\Pr(P|\sim C) &= 0.01 \\
\Pr(C) &= 0.01 \\
\Pr(\sim C) &= 0.99
\end{aligned}
$$

Base Rates Blues (continued)

Bayes' Rule thus means that:

$$
\begin{aligned}
\Pr(C|P) &= \frac{\Pr(P|C) \times \Pr(C)}{\Pr(P)} \\
&= \frac{\Pr(P|C) \times \Pr(C)}{[\Pr(P|C) \times \Pr(C)] + [\Pr(P|\sim C) \times \Pr(\sim C)]} \\
&= \frac{1.0 \times 0.01}{[1 \times 0.01] + [0.01 \times 0.99]} \\
&= \frac{0.01}{0.01 + 0.0099} \\
&= \mathbf{0.5025}
\end{aligned}
$$

Bayes' Rule Applied

For our data example above:

$$\begin{aligned}\Pr(\theta|X) &= \frac{\Pr(\theta \cap X)}{\Pr(X)} \\ &= \frac{\Pr(X|\theta)\Pr(\theta)}{\Pr(X)}.\end{aligned}$$

where:

- $\Pr(X|\theta)$ is the sampling density
- $\Pr(\theta)$ is the prior density of θ
- $\Pr(\theta|X)$ is the posterior density of θ
- $\Pr(X)$ is the marginal probability of X

Since X is fixed in a single sample, we can write:

$$\Pr(\theta|X) \propto \Pr(X|\theta)\Pr(\theta).$$

Bayes and Subjective Probability

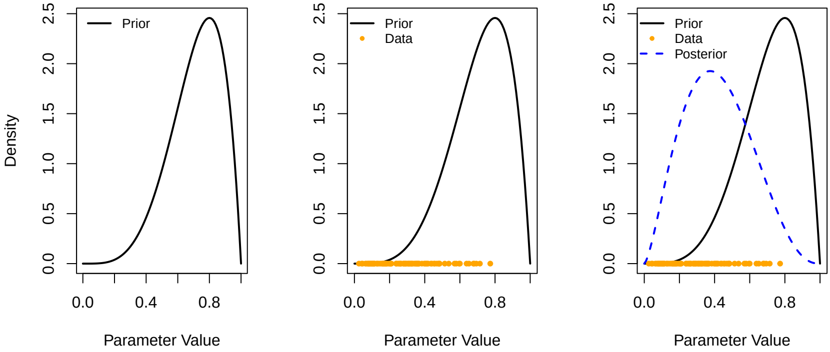
In the Bayesian view of probability:

- Probability is a belief about the world
  - Two individuals may/will have different beliefs about the probability of some event, *but*
  - Beliefs must still conform to the axioms, and be *rational*
  - Details are (e.g.) here

- $\Pr(\theta)$ is our prior / "pre-data" estimate of the value/distribution of θ

- $\Pr(\theta|X)$ is our posterior / "post-data" estimate

Bayes: Intuition



Bayesian Data Analysis

Steps:

- Posit a probability model for the data
- Posit one's prior beliefs
- Calculate the posterior distribution using Bayes' Theorem
- Summarize the posterior density
- Conduct post-estimation model checking

Bayesian Data Analysis (for real)

Parameter estimation (of, say, θ) essentially amounts to learning:

$$\mathsf{Pr}(\theta|X) = \frac{\mathsf{Pr}(X|\theta)\,\mathsf{Pr}(\theta)}{\mathsf{Pr}(X)}$$

The "evidence" part, $\mathsf{Pr}(X)$ (the "normalization factor") is, formally:

$$\mathsf{Pr}(X) = \int_{\theta} \mathsf{Pr}(X|\theta)\,\mathsf{Pr}(\theta)\,d\,\theta$$

For large numbers of parameters θ, this becomes analytically / computationally intractable. So...

Bayesian Data Analysis (for real)

"Now we might say 'OK, if we can't solve something, could we try to approximate it? For example, if we could somehow draw samples from that posterior we can Monte Carlo approximate it.' Unfortunately, to directly sample from that distribution you not only have to solve Bayes formula, but also invert it, so that's even harder.

Then we might say 'Well, instead let's construct an ergodic, reversible Markov chain that has as an equilibrium distribution which matches our posterior distribution.' I'm just kidding, most people wouldn't say that as it sounds bat-shit crazy. If you can't compute it, can't sample from it, then constructing that Markov chain with all these properties must be even harder.

The surprising insight though is that this is actually very easy and there exist a general class of algorithms that do this called Markov chain Monte Carlo (constructing a Markov chain to do Monte Carlo approximation)."

– Tom Wiecki, "MCMC Sampling for Dummies"

Markov Chain Monte Carlo (MCMC) methods:

- Sample from the posterior distribution $\Pr(\theta|X)$,
- using (multiple) autoregressive "markov chains," each starting at a different place in the parameter space.
- Those chains are essentially jumps around in the parameter space, according to a function of the likelihood at that point i ($\Pr(X|\theta_i)$).
- There's a *lot* more to this, some of which we'll return to in PLSC 503, and most of which you'll want to take a course in Bayesian statistics to learn...

Software

BUGS / WinBUGS

- Bayesian (Inference) Using Gibbs Sampling
- The OG of Bayesian software
- Still useful, but somewhat superceded
- Can be called from (e.g.) R (via `BRugs`, `R2WinBUGS`, `R2OpenBUGS`)

JAGS

- Just Another Gibbs Sampler
- Similar to BUGS...
- Can also be called from R (e.g., using `rjags`)

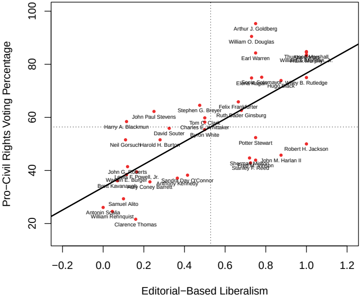
STAN

- Newer, faster, probably cooler
- Well beyond Gibbs sampling (MCMC, variational inference, penalized MLE)
- Interfaces with R via `rstan`, `rstanarm`, `brms`, and others

Example: SCOTUS Regression Re-Redux

You know the drill:

```
> describe(SCOTUS,skew=FALSE,trim=0)
                     vars  n    mean    sd  median     min     max range   se
justice                 1 39   97.87 11.60   98.00   78.00  117.00  39.0 1.86
justiceName*            2 39   20.00 11.40   20.00    1.00   39.00  38.0 1.83
CivLibs                 3 39   56.39 19.90   55.42   21.63   95.33  73.7 3.19
Nom.Order*              4 39   20.00 11.40   20.00    1.00   39.00  38.0 1.83
Nominee*                5 39   20.00 11.40   20.00    1.00   39.00  38.0 1.83
ChiefJustice*           6  4    1.00  0.00    1.00    1.00    1.00   0.0 0.00
SenateVote*             7 39   16.69  8.42   19.00    1.00   25.00  24.0 1.35
IdeologyScore           8 39    0.53  0.33    0.50    0.00    1.00   1.0 0.05
QualificationsScore*    9 39   16.38  7.82   18.00    1.00   25.00  24.0 1.25
Nominator (Party)*     10 39    6.92  3.72    6.00    1.00   13.00  12.0 0.60
Year                   11 39 1971.03 25.66 1967.00 1937.00 2020.00  83.0 4.11
```



Simple Bayesian Regression via brms

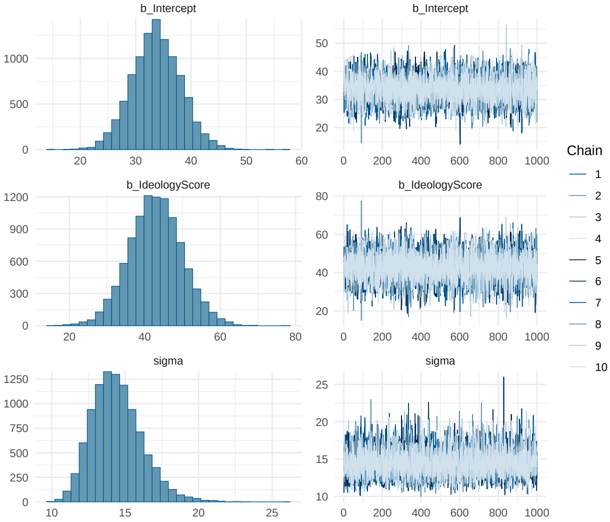
```
> # Default priors:
>
> get_prior(CivLibs~IdeologyScore,data=SCOTUS) # default prior...
                  prior     class          coef group resp dpar nlpar lb ub       source
                 (flat)         b                                                default
                 (flat)         b IdeologyScore                             (vectorized)
 student_t(3, 55.4, 25.2) Intercept                                              default
   student_t(3, 0, 25.2)     sigma                                  0            default
> bfit<-brm(CivLibs~IdeologyScore,data=SCOTUS,
+           chains=10,silent=2,seed=7222009)
.
.
.
> summary(bfit) # a la summary(lm())
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: CivLibs ~ IdeologyScore
   Data: SCOTUS (Number of observations: 39)
  Draws: 10 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 10000

Regression Coefficients:
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept        33.72      4.41    25.09    42.50 1.00    10184     7427
IdeologyScore    42.88      7.08    29.01    56.93 1.00    10207     6885

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma    14.44      1.71    11.56    18.23 1.00     8276     7056

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Post-Estimation Plots



Model Comparison

	OLS	Bayesian
(Intercept)	33.69^{*}	33.72^{*}
	(4.26)	[26.53; 40.91]
Ideology Score	42.94^{*}	42.88^{*}
	(6.85)	[31.46; 54.84]
R^2	0.52	0.50
Adj. R^2	0.50	
N	39	39
loo IC		320.50
WAIC		320.45

$^{*}p < 0.05$ (or Null hypothesis value outside the confidence interval).

Using Priors

Start with prior belief that $\beta_{\text{Ideology Score}} \sim \mathcal{N}(20, 225)$ (so s.d. $= 15$)...

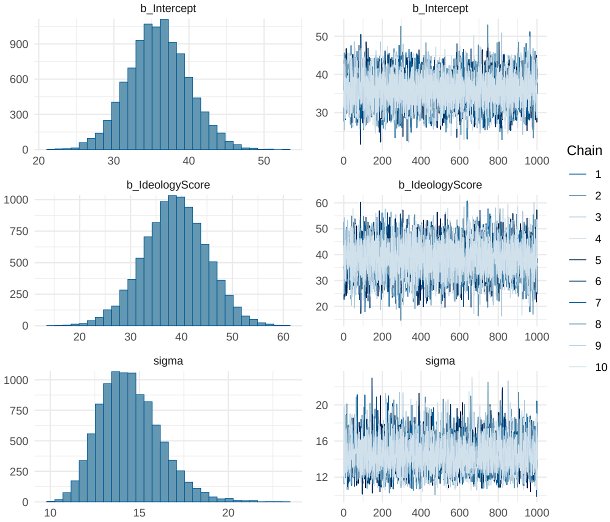
```
> Prior<-c(set_prior("normal(20,15)",class="b",coef="IdeologyScore"))
>
> bfit2<-brm(CivLibs~IdeologyScore,data=SCOTUS,chains=10,silent=2,seed=7222009,
+            prior=Prior)
.
.
.
> summary(bfit2) # a la summary(lm())
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: CivLibs ~ IdeologyScore
   Data: SCOTUS (Number of observations: 39)
  Draws: 10 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 10000

Regression Coefficients:
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept        35.92      4.01    28.14    44.02 1.00     9556     7151
IdeologyScore    38.73      6.26    26.01    50.82 1.00     9634     7285

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma    14.47      1.72    11.58    18.25 1.00     7749     6475

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Post-Estimation Plots

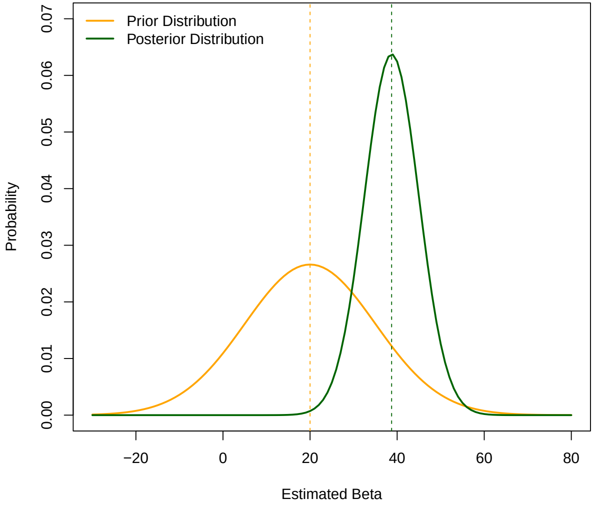


Model Comparison (again)

	OLS	Bayesian	w/Priors
(Intercept)	33.69^*	33.72^*	35.92^*
	(4.26)	[26.53; 40.91]	[29.40; 42.61]
Ideology Score	42.94^*	42.88^*	38.73^*
	(6.85)	[31.46; 54.84]	[28.32; 48.83]
R^2	0.52	0.50	0.45
Adj. R^2	0.50		
N	39	39	39
loo IC		320.50	320.43
WAIC		320.45	320.39

$^*p < 0.05$ (or Null hypothesis value outside the confidence interval).

Prior vs. Posterior for $\beta_{\text{Ideology Score}}$



Bayes: Pros

- **Directly quantifies uncertainty**
- **Provides direct quantities of interest to researchers**
- **Logically consistent and intuitive**
- **Allow the incorporation of prior information**
- **Allow the fitting complex models**
- **Flexibility**

Bayes: Cons

- **Inherent subjectivity of choosing priors**
- **Computational complexity**
- **Difficulty in knowing when estimates have converged**
- **Sensitivity to analyst choices** (but see, e.g., the "WAMBS checklist")
- **Slow uptake among applied social scientists**