Fully Bayesian Models

Bayesian Idea: Treat a parameter as having an unknown **Distribution** to be estimated, rather than as an unknown fixed number to be estimated.

Ex 1 Predict end-of-season win percentage from mid-season Wins and Losses.

Beta-Binomial model

$$\begin{cases} W \sim \text{Binomial}(n, p) \\ p \sim \text{Beta}(\alpha, \beta) \end{cases}$$

When we model the latent team's win probability p using a Beta Prior, we encode the prior information that p is more likely to be near $1/2$ (say, in $.3, .7$) than to be very near 0 or 1.

Then, we found using Bayes Rule that the posterior distribution $p|W,L$ is

$$p|W,L \sim \text{Binomial}\left(n+\alpha+\beta-2, \frac{W+\alpha-1}{W+L+\alpha+\beta-2}\right)$$

and the Bayes estimate is the posterior mean

$$\hat{p}^{(\text{Bayes})} = \mathbb{E}[p|W,L] = \frac{W+(\alpha-1)}{W+L+(\alpha+\beta-2)}.$$

Ex 2 Predict end-of-season batting average from mid-season batting average and number of at-bats.

Normal-Normal model

i = player i

H_i = # hits, N_i = # at-bats, $X_i = \frac{H_i}{N_i}$

$X_i \sim N(\mu_i, \sigma_i^2)$

$\sigma_i^2 = C/N_i$, C known

$\mu_i \sim N(\mu, \tau^2)$

When we model player i's latent quality μ_i using a Normal Prior, we encode the prior information that player i is also a baseball player, allowing us to share strength across players.

Then, we found using Bayes Rule that the posterior distribution $\mu_i | X$ is

$$\mu_i | X \sim N\left(\frac{\left(\frac{X_i}{\sigma_i^2} + \frac{\mu}{\tau^2}\right)}{\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)}, \frac{1}{\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)} \right)$$

and the Bayes estimate is the posterior mean

$$\hat{\mu}_i^{(\text{Bayes})} = \frac{\frac{X_i}{\sigma_i^2} + \frac{\hat{\mu}}{\hat{\tau}^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\hat{\tau}^2}}$$

using $\hat{\mu}$ and $\hat{\tau}^2$ since μ, τ are unknown.

Ex 3 Bayesian Regression:

model $\begin{cases} \text{regression} & y_i \sim N(x_i^T\beta, \sigma^2) \\ \text{prior} & \beta \sim N(0, \frac{\sigma^2}{\lambda}) \end{cases}$

If you use Bayes Rule to find the posterior distribution $P(\beta | X, y)$, you'll find

$$\beta | X, y \sim N(\underbrace{(X^TX + \lambda I)^{-1}X^Ty}_{\text{Ridge Regression}}, \text{something})$$

* To make decisions in sports (e.g. player valuation or play selection), we need not only know our best estimate of the value of the player/play/decision, but also uncertainty quantification (e.g., error bars or prediction intervals) to describe how confident/certain we are about a value.

Bayesian Idea Estimate the FULL Posterior Distribution of an unknown parameter. This gets us error bars (uncertainty quantification), with which we can create more complete decisions.

Q. Create a fully Bayesian NFL power Rating model.

(Glickman Stern 1998).

Initial Dataframe:

game_id	home_team	away_team	season_type	week	total_home_score	total_away_score	season	pts_H_minus_A
2018_01_ATL_PHI	PHI	ATL	REG	1	18	12	2018	6
2018_01_BUF_BAL	BAL	BUF	REG	1	47	3	2018	44
2018_01_CHI_GB	GB	CHI	REG	1	24	23	2018	1
2018_01_CIN_IND	IND	CIN	REG	1	23	34	2018	-11
2018_01_DAL_CAR	CAR	DAL	REG	1	16	8	2018	8
2018_01_HOU_NE	NE	HOU	REG	1	27	20	2018	7
2018_01_JAX_NYG	NYG	JAX	REG	1	15	20	2018	-5
2018_01_KC_LAC	LAC	KC	REG	1	28	38	2018	-10
2018_01_LA_OAK	LV	LA	REG	1	13	33	2018	-20
2018_01_NYJ_DET	DET	NYJ	REG	1	17	48	2018	-31
2018_01_PIT_CLE	CLE	PIT	REG	1	21	21	2018	0
2018_01_SEA_DEN	DEN	SEA	REG	1	27	24	2018	3

Create variables that make it easier to write the model:

```
# A tibble: 1,657 x 13
   game_id         home_team away_team season_type  week total_home_score total_away_score season pts_H_minus_A     S     H     A     y
   <chr>           <chr>     <chr>     <chr>       <dbl>            <dbl>            <dbl>  <dbl>          <dbl> <dbl> <int> <int> <dbl>
 1 2018_01_ATL_PHI PHI       ATL       REG             1               18               12   2018              6     1    26     2     6
 2 2018_01_BUF_BAL BAL       BUF       REG             1               47                3   2018             44     1     3     4    44
 3 2018_01_CHI_GB  GB        CHI       REG             1               24               23   2018              1     1    12     6     1
 4 2018_01_CIN_IND IND       CIN       REG             1               23               34   2018            -11     1    14     7   -11
 5 2018_01_DAL_CAR CAR       DAL       REG             1               16                8   2018              8     1     5     9     8
 6 2018_01_HOU_NE  NE        HOU       REG             1               27               20   2018              7     1    22    13     7
 7 2018_01_JAX_NYG NYG       JAX       REG             1               15               20   2018             -5     1    24    15    -5
 8 2018_01_KC_LAC  LAC       KC        REG             1               28               38   2018            -10     1    18    16   -10
 9 2018_01_LA_OAK  LV        LA        REG             1               13               33   2018            -20     1    19    17   -20
10 2018_01_NYJ_DET DET       NYJ       REG             1               17               48   2018            -31     1    11    25   -31
# i 1,647 more rows
```

Variables

Row i = game i in the dataset

$H(i)$ = index of the home team in the i^{th} game

$A(i)$ = away team

$S(i)$ = index of the season (year) in the i^{th} game

y_i = pts scored by $H(i)$ − pts scored by $A(i)$

Model

$$y_i \sim N\left(\beta_0 + \beta_{H(i),S(i)} - \beta_{A(i),S(i)},\ \sigma^2_{games}\right)$$

Bayesian Model: All parameters have distributions

Autoregressive prior

$$\beta_{j,s} \sim N\left(\gamma \cdot \beta_{j,s-1},\ \sigma^2_{seasons}\right)$$

for all teams j and seasons $s > 1$

Priors

$\beta_{j,1} \sim N(0, \sigma^2_{teams})$ $\quad \beta_0 \sim N(0, 5^2)$

$\sigma^2_{teams} \sim N_+(0, 5^2)$ $\quad \sigma^2_{seasons} \sim N_+(0, 5^2)$

$\sigma^2_{games} \sim N_+(0, 5^2)$ $\quad \gamma \sim Unif[0, 1]$

* In a general Bayesian statistical model, the parameters don't all have Gaussian priors or likelihoods, or the model is quite large and complicated, so we can't write on paper a closed-form analytical solution for the posterior distribution.

* Typically need to approximate the posterior distribution using MCMC sampling methods like
  - Gibbs sampling
  - Hamiltonian Monte Carlo
  - No U-Turn Sampling,

which we won't cover here (take Shane's class), using a probabilistic programming language
- Stan
- Jags
- NumPyro

Fit a fully Bayesian model using Stan

in a file called "bayesian_model_glickmanStern.stan":

```
data {
  int<lower=1> N_games;            // number of games
  int<lower=1> N_teams;            // number of teams
  int<lower=2> N_seasons;          // number of seasons

  real y[N_games];                 // outcome vector (point differential)
  int<lower=1, upper=N_teams> H[N_games];   // vector of home team indices
  int<lower=1, upper=N_teams> A[N_games];   // vector of away team indices
  int<lower=1, upper=N_seasons> S[N_games]; // vector of season indices
}
```

```
parameters {
  real beta_0;                    // intercept (home field advantage)
  real betas[N_teams, N_seasons];  // team strength coefficients for each team-season

  real<lower=0> sigma_games;      // game-level variance in point differential
  real<lower=0> sigma_teams;      // variance across teams before the first season
  real<lower=0> sigma_seasons;    // a team's variance across seeasons
  real<lower=0, upper=1> gamma;   // autoregressive parameter
}
```

```
model {
  // game-level model
  for (i in 1:N_games) {
    y[i] ~ normal(beta_0 + betas[H[i],S[i]] - betas[A[i],S[i]], sigma_games);
  }

  // team-level priors
  for (j in 1:N_teams) {
    // initial season prior across teams
    betas[j,1] ~ normal(0, sigma_teams);
    for (s in 2:N_seasons) {
      // auto-regressive model across seasons
      betas[j,s] ~ normal(gamma*betas[j,s-1], sigma_seasons);
    }
  }

  // priors
  sigma_games ~ normal(0, 5);
  sigma_teams ~ normal(0, 5);
  sigma_seasons ~ normal(0, 5);
  gamma ~ uniform(0, 1);
}
```

* Stan uses MCMC Hamiltonian Monte Carlo to approximate the posterior distribution of each parameter.
Stan returns the approximate posterior dist via posterior samples.
For example, the fitted posterior of β_{js} is m draws $\beta_{js}^{(1)}, \ldots, \beta_{js}^{(m)}$ from the approximate posterior.

Fit the model from R using RStan:

```
library(rstan)
```

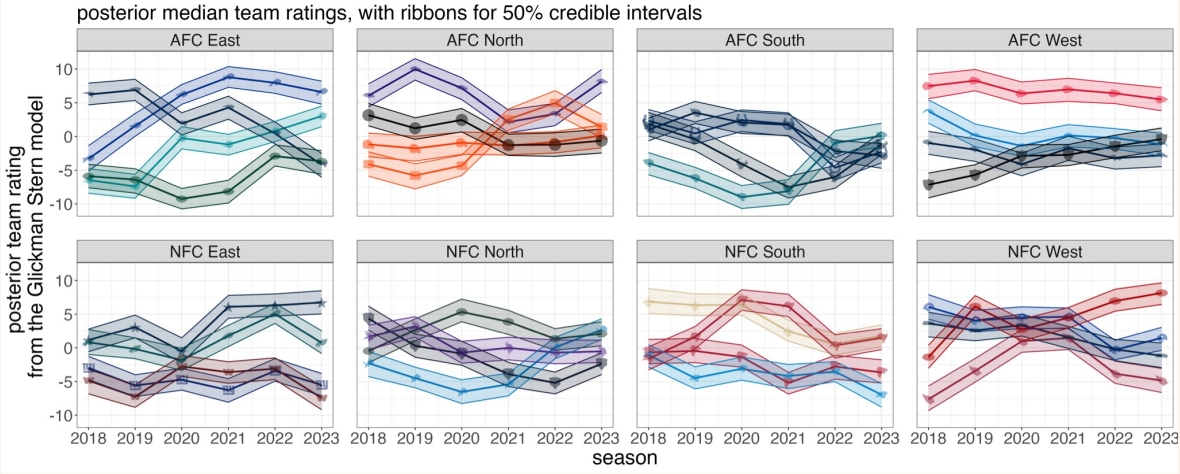
```
### load stan model
MODEL <- stan_model(file = "bayesian_model_glickmanStern.stan", model_name = "glickmanSternModel")
MODEL

### create list of data compliant with the Stan model
data_train <- list(
  N_games = nrow(df1),
  N_teams = nrow(map_team_to_idx),
  N_seasons = length(unique(df1$season)),
  y = df1$y,
  H = df1$H,
  A = df1$A,
  S = df1$S
)
data_train

# Train the model
fit <- sampling(
  MODEL, data = data_train, iter = 1500, chains = 1, seed = 12345,
)
fit
```

Results

	param	post_lower	post_med	post_upper
	<chr>	<dbl>	<dbl>	<dbl>
1	beta_0	0.968	1.53	2.08
2	sigma_games	12.0	12.4	12.9
3	sigma_teams	3.70	4.94	6.75
4	sigma_seasons	3.02	3.70	4.49
5	gamma	0.493	0.662	0.814



There are many great sports papers that use fully Bayesian models because they

- are interpretable
- quantify uncertainty
- capture various sources of uncertainty or variance
- use shrinkage

Some examples:

- glickman and stern 1998
- Sameer & Wyner pitch framing
- Shane & Wyner fielding
- Brill & Wyner time through the order
- Lopez et al Home field advantage
- Lopez et al Comparing Randomness Across Sports