Getting Beyond the Mean in Predictive Inference

Hal Stern

Department of Statistics
University of California, Irvine

Conference on Predictive Inference and Its Applications
Iowa State University, Ames, IA

Predictive Inference

- Predictive Inference .... Statistical inference with the focus on observables rather than parameters
  - Geisser 1993 "Predictive Inference" book
  - Harville 2014 "The Need for More Emphasis on Prediction" (The American Statistician)
- Currently a great deal of focus on prediction
  - the rise of machine learning / artificial intelligence
  - focus on $\hat{Y}$ and generally not on $\hat{\theta}$
  - Breiman 2001 on "The Two Cultures" (Statistical Science)
- This talk ... a focus on model-based prediction in several application areas

College Football Ratings

DUNKEL INDEX

Sports Picks & Rankings Since 1929

College Football Ranking Composite

Monday, January 8, 2018 (103 Rankings)

| Last Week | Archived | FBS | FCS | BCS | CSV Data | Kaggle Archive | About | Schedule | Meta
Analysis: | Kislanko | Coleman | Wobus | Fremeau | Perez | Rankings | Superlist |

Payne, PiRate, OSCAR, Laz Index, RoundTable

ARGH, J-Train, Howell, Coffey, Henderson

Kirkpatrick, Kellner, Hatch, Adjusted Stats EWP, MvG Sports

Round Ro, Bihl, Junky's, Ashburn, Singer

PAY	PIR	OSC	LAZ	RT	ARG	JTR	HOW	COF	HEN	Rank	Team	Conf	Record	KPK	KEL	HAT	HKB	MvG	RWP	BIH
1	1	1	2	1	1	2	1	1	2	1	Alabama	SEC	13-1	2	1	1	1	1	2	1
4	2	3	1	4	4	5	3	6	5	2	Georgia	SEC	13-2	3	2	4	2	2	3	3
2	3	2	3	2	5	3	4	2	1	3	Ohio St	B10	12-2	1	4	6	5	6	6	5
3	4	4	4	5	2	1	2	3	3	4	Wisconsin	B10	13-1	4	3	3	3	5	1	4
6	7	5	5	6	3	6	5	5	7	5	Clemson	ACC	12-2	6	5	5	4	3	4	6
5	6	6	6	3	6	4	6	4	4	6	Penn St	B10	11-2	5	6	7	7	8	5	8
8	5	8	8	7	7	7	7	7	8	7	UCF	AAC	13-0	7	7	2	8	4	7	2
7	8	7	7	9	9	9	8	8	6	8	Oklahoma	B12	12-2	9	9	8	6	7	8	7
9	9	10	9	8	8	8	9	9	9	9	Notre Dame	FBSI	10-3	8	8	9	11	9	10	9
10	10	9	10	10	10	10	10	12	11	10	Auburn	SEC	10-4	10	13	15	9	12	9	10

Settings for Predictive Inference

- Rating sports teams
  - Linear / least-squares approach

    $$Y_{ij} = H + \theta_i - \theta_j + \epsilon_{ij}$$

    where
    Y_{ij} is game outcome
    θ_i is "strength" of team i
    H is home-field advantage
    ϵ_{ij} is variation/error
  - Many possible additions to the model
    (separate off./def. strengths, time-varying parameters, etc.)
  - Interested in ratings and predictions implied by ratings
  - This basic model does remarkably well
    (see, e.g., Harville, 1977, 1980; Stern, 1995; many others)

Settings for Predictive Inference

- Animal Breeding
  - Mixed linear model

$$Y = X\beta + Z\theta + \epsilon$$

    where
    Y $(n \times 1)$ is a vector of measures of phenotypic trait of interest
    $X\beta$ is contribution from fixed effects (e.g., gender)
    $Z\theta$ is contribution from random effects (e.g., genetic effects, shared environment)
  - Interested in inference for parameters with a focus on predicting quality of future generations

Settings for Predictive Inference

- Disease Mapping (small area estimation)
  - Poisson hierarchical model

$$Y_i \sim \theta_i E_i$$

$$\lambda = log(\theta) \sim N(X\beta, V(\sigma))$$

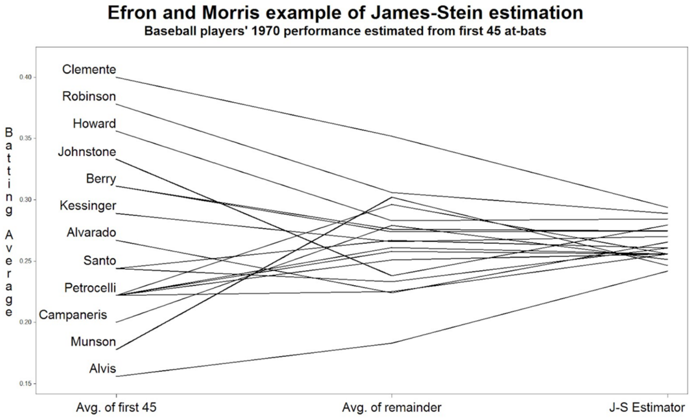
    where
    Y are observed disease incidence counts,
    E expected counts based on demographics,
    θ are parameters measuring "risk",
    $X\beta$ measures contribution of covariates to risk,
    σ are parameters of the variance matrix
  - Interested in estimates of θ
    (and their implied predictions of future incidence)

Settings for Predictive Inference

- The three problems share a common statistical structure in which observed outcomes are modelled in terms of parameters including a set of unit-level parameters θ
  - θ is linked to predictions of future observables
  - Traditionally viewed as random effects
  - Bayesian view is as part of a hierarchical model

Inference through the (posterior) mean

- Under squared-error loss optimal estimator of θ is the posterior mean ($\hat{\theta} = E(\theta|y)$)
- James-Stein estimation example from Efron and Morris (1977, Scientific American)



Inference through the (posterior) mean

- There are some issues associated with relying on the posterior mean
  - Louis (1984), Ghosh (1992) show that

    $$Var(E(\theta|Y)) < E(Var(\theta|Y))$$

    so the posterior mean estimates "shrink too much" in that their variance doesn't reflect the variation in the ensemble
  - Ignores $Var(\theta_i|Y)$ which may exhibit substantial variation
  - Scientific problem may suggest another loss function

Getting beyond the (posterior) mean

- Approaches to prediction that don't rely only on the posterior mean
  - Alternative (richer) model parameterizations
  - Alternative loss functions
  - Changing the estimand (alternative posterior inferences)

Getting beyond the mean - alternative parameterizations

- Example - Chess ratings
  - Traditional paired comparison model introduces a strength parameter for each player (θ_i)
  - $Pr(i > j) = \frac{10^{(\theta_i - \theta_j)/400}}{1+10^{(\theta_i - \theta_j)/400}}$
  - Ties can be accommodated but not addressed here
  - Traditional (Elo) approach updates after each game but
    - This ignores uncertainty about ratings
    - Does not deal appropriately with changes in ability over time

Chess ratings

- Glickman (1993, 1999) introduces an enhanced model
  - Each player is characterized by two parameters, $\theta_i \sim N(\mu_i, \sigma_i^2)$
  - Ratings updated based on results in a rating period (binomial (or trinomial) likelihood as above)
  - Inferences obtained by averaging over uncertainty in θ
  - Players μ_i and σ_i^2 are updated
  - Variances increase due to passage of time (without playing)

Chess ratings

Rank	*Player*	*Posterior mean strength in peak year*	*Posterior standard deviation*	*Peak year*
1	Emanuel Lasker	1693	29	1916
2	José Capablanca	1680	28	1921
3	Robert Fischer	1656	38	1972
4	Alexander Alekhine	1647	24	1930
5	Garry Kasparov	1643	32	1991
6	Mikhail Botvinnik	1623	27	1947
7	Anatoly Karpov	1609	20	1984
8	Wilhelm Steinitz	1608	29	1876
9	Akiba Rubinstein	1584	24	1912
10	Max Euwe	1579	23	1935
11	Boris Spassky	1578	21	1968
12	Siegbert Tarrasch	1576	25	1905
13	Viktor Korchnoi	1573	21	1978
14	Geza Maroczy	1572	25	1908
15	David Bronstein	1571	23	1953
16	Vassily Ivanchuk	1570	32	1991
17	Samuel Reshevsky	1569	24	1952
18	Vassily Smyslov	1567	22	1954
19	Aron Nimzovitch	1565	26	1931
20	Tigran Petrosian	1564	22	1963

Getting beyond the mean - alternative loss functions

- Example - Disease mapping
  - Recall our Poisson hierarchical model

$$Y_i \sim \theta_i E_i$$

$$\lambda = log(\theta) \sim N(X\beta, V(\sigma))$$

    where
    Y are observed counts,
    E expected counts based on demographics,
    θ are parameters measuring "risk",
    $X\beta$ measures contribution of covariates to risk,
    σ are parameters of the variance matrix
  - Common to choose estimates to minimize expected sum of squared error loss (SSEL)

$$\sum_{k=1}^{n} (\theta_k - \hat{\theta}_k)^2$$

    which leads to $\hat{\theta}_k = E(\theta_k | Y)$
  - But we know these estimates are underdispersed ... and we are often interested in extrema

Getting beyond the mean - alternative loss functions

- Example - Disease mapping
  - Wright et al., (2003) introduce weighted-ranks squared error loss (WRSEL)

    $$WRSEL_c(\theta, \hat{\theta}) = \sum_{k=1}^{n} c_{r(k)}(\theta_k - \hat{\theta}_k)^2$$

    where $r(k)$ =rank of θ_k among the K regions
    and c is a vector of weights identifying inferential priorities
  - Estimate is a weighted average of conditional posterior means

$$\hat{\theta}_k = \frac{\sum_{j=1}^{n} c_j Pr(\theta_k = \theta_{(j)}|Y)E(\theta_k|\theta_k = \theta_{(j)}, Y)}{\sum_{j=1}^{n} c_j Pr(\theta_k = \theta_{(j)}|Y)}$$

Scotland Lip Cancer Data

TABLE 1: Scotland-lip-cancer data.

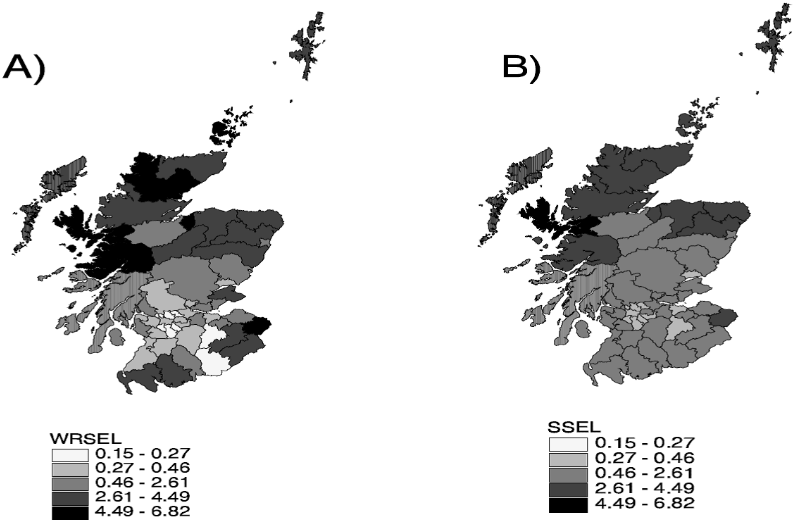
ID	District	y	E	y/E	AFF	Neighbours
1	Skye-Lochalsh	9	1.38	6.52	16	5, 9, 11, 19
2	Banff-Buchan	39	8.66	4.50	16	7, 10
3	Caithness	11	3.04	3.62	10	6, 12
4	Berwickshire	9	2.53	3.56	24	18, 20, 28
5	Ross-Cromarty	15	4.26	3.52	10	1, 11, 12, 13, 19
6	Orkney	8	2.40	3.33	24	3, 8
7	Moray	26	8.11	3.21	10	2, 10, 13, 16, 17
8	Shetland	7	2.30	3.04	7	6
9	Lochaber	6	1.98	3.03	7	1, 11, 17, 19, 23, 29
10	Gorden	20	6.63	3.02	16	2, 7, 16, 22
11	WesternIsles	13	4.40	2.95	7	1, 5, 9, 12
12	Sutherland	5	1.79	2.79	16	3, 5, 11
13	Nairn	3	1.08	2.78	10	5, 7, 17, 19
14	Wigtown	8	3.31	2.42	24	31, 32, 35
15	NEFife	17	7.84	2.17	7	25, 29, 50
16	Kincardine	9	4.55	1.98	16	7, 10, 17, 21, 22, 29
17	Badenoch	2	1.07	1.87	10	7, 9, 13, 16, 19, 29
18	Ettrick	7	4.18	1.67	7	4, 20, 28, 33, 55, 56
19	Inverness	9	5.53	1.63	7	1, 5, 9, 13, 17
20	Roxburgh	7	4.44	1.58	10	4, 18, 55
21	Angus	16	10.46	1.53	7	16, 29, 50
22	Aberdeen	31	22.67	1.37	16	10, 16

Scotland Lip Cancer Data

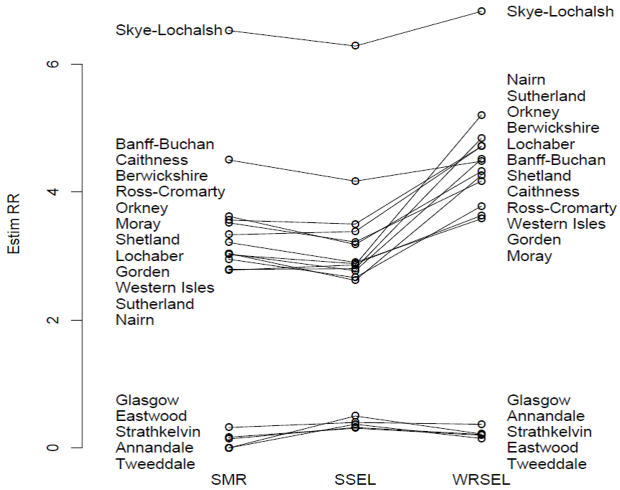
- Apply WRSEL (and compare to SSEL)
- Use c to be "bowl-shape"
  (large weight on highest and lowest order statistics)

District	SMR	SMR rank	WRSEL estimate	WRSEL rank	SSEL estimate	SSEL rank
Skye-Lochalsh	6.52	56	6.82	56	6.28	56
Banff-Buchan	4.50	55	4.48	50	4.17	55
Caithness	3.62	54	4.32	47	3.18	51
Berwickshire	3.56	53	4.71	52	3.50	54
Ross-Cromarty	3.52	52	4.17	56	3.22	52
Orkney	3.33	51	4.71	53	3.38	53
Moray	3.21	50	3.58	42	2.90	50
Shetland	3.04	49	4.36	48	2.62	44
Lochaber	3.03	48	4.51	51	2.77	46
Gorden	3.02	47	3.63	43	2.88	49
WesternIsles	2.95	46	3.78	34	2.66	45
Sutherland	2.79	45	4.34	54	2.80	47
Nairn	2.78	44	5.20	55	2.86	48
⋮						⋮
Glasgow	0.32	8	0.37	18	0.40	7
⋮						⋮
Eastwood	0.17	4	0.20	4	0.31	1
Strathkelvin	0.14	3	0.21	5	0.32	2
Annandale	0.00	2	0.22	6	0.50	12
Tweeddale	0.00	1	0.15	1	0.37	4

Scotland Lip Cancer Data



Scotland Lip Cancer Data



Getting beyond the mean - alternative loss functions

- Triple goal estimates (Shen and Louis, 1998) - estimation procedure that simultaneously targets
  - estimates for each individual unit
  - estimation of the distribution of parameters
  - estimation of the ranks of the units
- Quantile estimates (Ginestet, 2011 thesis) - squared error loss on a set of specified quantiles of the ensemble distribution

Getting beyond the mean - alternative posterior summaries

- With a Bayesian analysis, we often approximate the posterior distribution $p(\theta|Y)$ via simulation
- Have simulations $\theta^{(s)}, s = 1, \ldots, S$ from $p(\theta|Y)$
- Can use these simulations to summarize any characteristic of the posterior distribution
  - posterior mean
  - WRSEL estimates
  - distribution of rank of parameter θ_k
  - $Pr(\theta_k > M)$ where M is a relevant risk factor
  - $Pr(R_k \leq 10)$ where R_k is rank of θ_k

Infant mortality rates

- Project with NCHS examining county-level infant mortality (death within the first year) rates
- n_i is number of births and Y_i number of deaths in county i during the years 1994-1998
- Statistical model

$$Y_i \sim Binom(n_i, \theta_i), i = 1, \ldots, I$$

$$logit(\theta_i) \sim N(\mu, \tau^2)$$

- Project also considered alternative models taking into account geographical relationships (health service areas, states, regions, etc.)
- Results shown for $1000\theta_i$, i.e., deaths per 1000

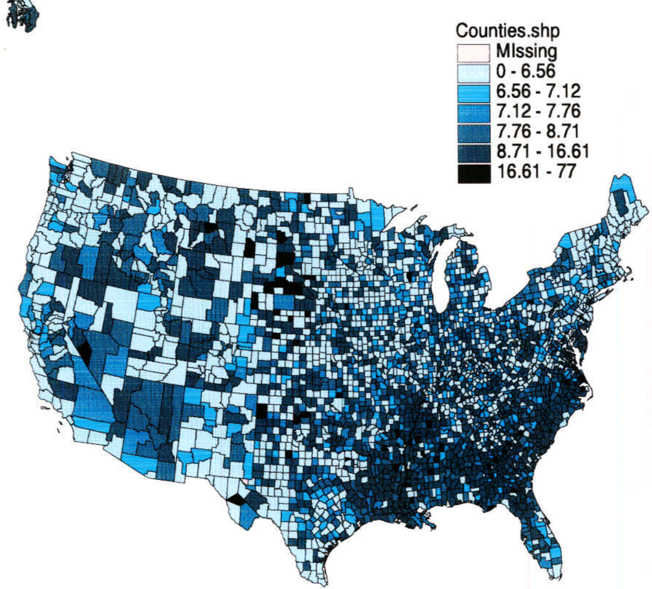


Figure 1: Observed county-level IMR's.

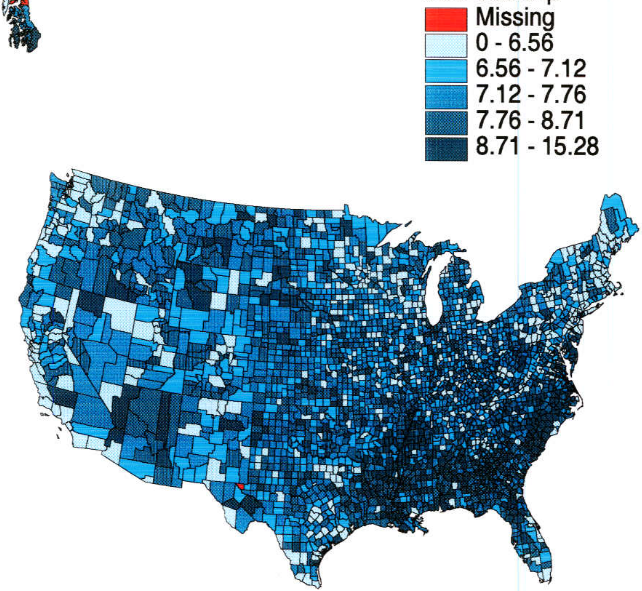


Figure 2: Estimated IMR's from the national model.

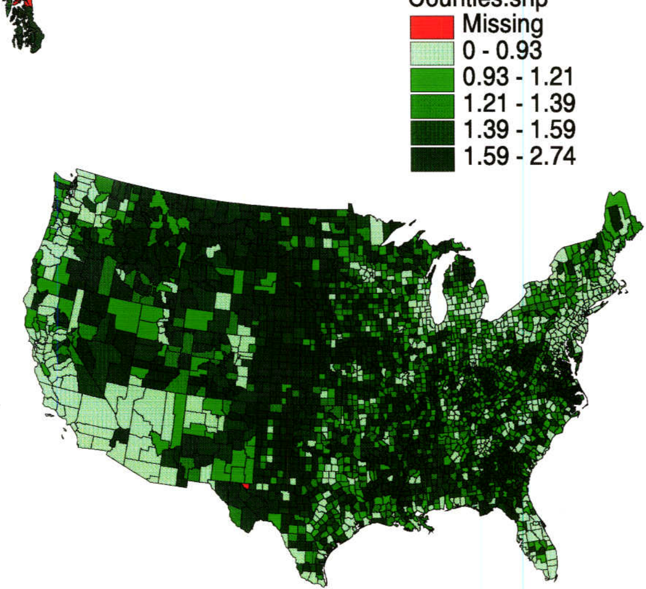


Figure 3: Posterior standard deviation of the IMR's from the national model.

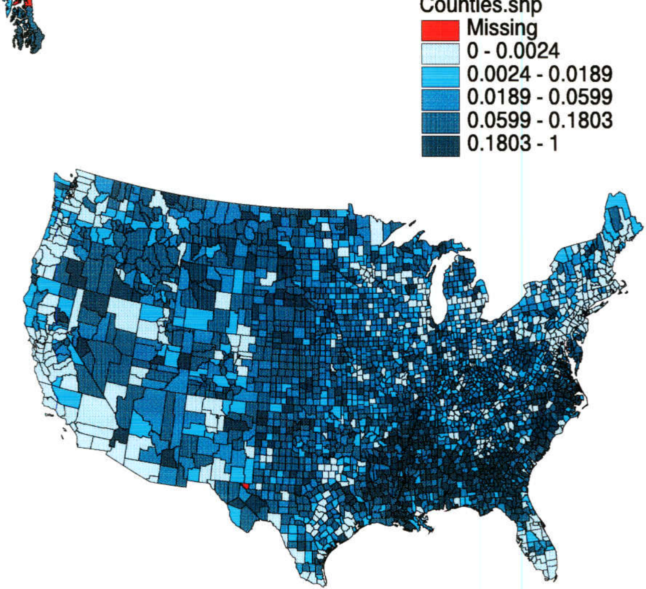


Figure 4: Posterior probabilities that IMR > 10 from the national model.

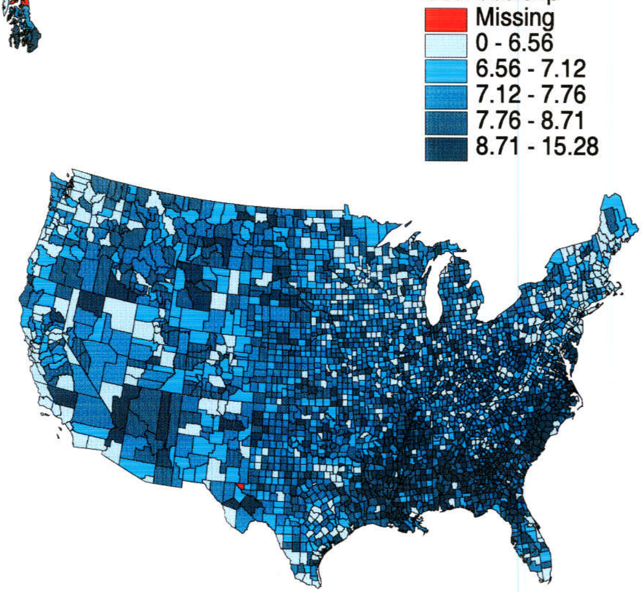


Figure 2: Estimated IMR's from the national model.

Infant mortality rates

County	Births	Obs	Est	SD	Pr(>10)
Kings, NY	206060	9.2	9.2	0.2	0.00
DeKalb, GA	49024	9.3	9.2	0.4	0.03
Potter, TX	10329	9.5	9.2	0.8	0.16
Rockingham, NC	5749	9.7	9.2	1.1	0.22
Jasper, IN	1802	11.1	9.2	1.6	0.29
Mellette, SD	200	25.0	9.2	2.2	0.32
La Salle, TX	432	16.2	9.2	2.1	0.33

Animal breeding selection experiments

- The mixed linear model ($Y = X\beta + Z\theta + \epsilon$) has been commonly used in analysis of animal breeding selection experiments
- Estimates of θ (genetic effects) are used to select animals for breeding
- Traditional inference (Henderson et al. 1959, Harville 1974)
  - estimate variance components with restricted maximum likelihood (REML)
  - estimate/predict the breeding values (θ's) using best-linear unbiased prediction (BLUP)
  - these are also the posterior means of the conditional on the variance components
- Bayesian inference is also now popular which averages over the posterior distribution of the variance components (Gianola and Fernando 1986)

Animal breeding selection experiments

- Can rank by posterior mean ... but as in the infant mortality data may be better to take uncertainty into account

		Estimated probability of being in		
Animal	*Posterior mean*	*Top 1*	*Top 3*	*Top 10*
146	116	0.042	0.108	0.237
12	94	0.014	0.046	0.144
148	104	0.025	0.077	0.191
147	94	0.022	0.056	0.160
149	95	0.022	0.061	0.169
127	105	0.035	0.084	0.203
252	91	0.025	0.061	0.155
248	72	0.008	0.029	0.098
377	73	0.010	0.030	0.099
250	69	0.007	0.027	0.087
375	67	0.010	0.026	0.080
251	66	0.006	0.026	0.085
253	65	0.009	0.022	0.079
126	81	0.013	0.041	0.118
374	64	0.008	0.023	0.078
304	89	0.020	0.057	0.144
379	54	0.007	0.020	0.064
151	52	0.007	0.018	0.063
249	54	0.005	0.018	0.061
34	60	0.005	0.018	0.068
86	78	0.014	0.039	0.117
254	80	0.016	0.043	0.118
85	84	0.024	0.055	0.140
315	95	0.034	0.074	0.173
178	70	0.012	0.036	0.099

Getting beyond the mean in predictive inference

- Predictive inference has proven valuable in many settings
- Likely to be increasingly important in the future (e.g., personalized / precision medicine)
- Important to consider relevant predictive summaries
  - this includes more than the population mean (e.g., treatment heterogeneity)
  - it also includes looking at more than posterior means of unit-level parameters
  - account for uncertainty in unit-level parameters
  - defining problem-specific esitimands or loss functions
- The use of probabilistic summaries also impacts model evaluation - need to make sure inferences are well-calibrated
- Questions/comments: sternh@uci.edu