SAS® GLOBAL FORUM 2020

MARCH 29 – APRIL 1

USERS PROGRAM

From Posterior to Postprocessing: Getting More from Your Bayesian Model

Matt Simpson, SAS Institute Inc.

- Originally from Missouri
- Iowa State University PhD in Stat + Econ
- University of Missouri postdoc in Stat
- Token Bayesian developer in SAS Econometrics (1.5 years)
- Default presentation template user
- Does not know where his prior comes from

The Bayesian Workflow

Step 0: Figure out the model and prior

Step 1: Fit the model

Step 2: Return to Step 0
unless you are satisfied

Step 3: Now what?

Bayesian Postprocessing

PROC QLIM and PROC COUNTREG give you a sample from the posterior distribution:

$$p(\theta|D) = \frac{p(D|\theta)\, p(\theta)}{\int p(D|\theta)\, p(\theta)\, \mathrm{d}\theta}$$

What if you don't care about θ directly?

Bayesian Postprocessing

Suppose you have a posterior sample: $\theta^{\{1\}}, \theta^{\{2\}}, \dots, \theta^{\{N\}}$

- $f\left(\theta^{\{1\}}\right), f\left(\theta^{\{2\}}\right), \dots, f\left(\theta^{\{3\}}\right)$ is a sample for $f(\theta)$

Posterior predictive distribution:

distribution of new observations, given the data

Posterior Predictive Distribution

$$p(D_{\text{new}}|D) = \int p(D_{\text{new}}|\theta)p(\theta|D)\mathrm{d}\theta$$

Obtain by sampling conditional on posterior draws:

$$D_{\text{new}}^{\{1\}} \sim p\left(D_{\text{new}} \middle| \theta^{\{1\}}\right)$$

$$D_{\text{new}}^{\{2\}} \sim p\left(D_{\text{new}} \middle| \theta^{\{2\}}\right)$$

$$\vdots$$

Example: 4x4 Truck Sales

Network of 100 dealerships

Want to predict potential new dealerships' sales using:

- Price
- Climate variables
- Demographic variables

Count (Poisson) regression

area_type	N Obs
rural	22
sub	52
urban	26

Variable	N	Mean	Std Dev	Minimum	Maximum
pop_bachelors	100	12118.38	2956.35	6684.00	20223.00
pop_below_bachelors	100	37857.67	2969.18	29703.00	43258.00
median_income	100	44012.19	13115.21	18261.00	80122.00
cost_of_living	100	127.36	20.26	78.00	176.00
mean_summer_temp	100	84.70	5.16	71.00	95.00
mean_winter_temp	100	34.19	8.18	11.00	60.00
mean_precip	100	23.83	12.21	5.00	92.00
price	100	25020.00	952.72	22600.00	27500.00
sales	100	177.60	123.79	48.00	469.00

Which Potential Location Is Better?

Obs	pop_bachelors	pop_below_bachelors	median_income	price	cost_of_living	mean_summer_temp	mean_winter_temp	mean_precip	area_type	sales
1	9760	40391	49683	25000	142	84	26	22	rural	.
2	13275	36545	52167	25000	124	79	51	24	urban	.
3	10431	39528	43897	25000	132	86	41	56	sub	.
4	7458	42359	49461	25000	110	88	35	33	sub	.
5	13038	36930	30133	25000	183	86	37	54	urban	.

Poisson Regression in PROC COUNTREG

```
proc countreg data = truckcount_transformed plots = none;
    class area_type;
    model sales = area_type log_pop_bachelors log_pop_below_bachelors
        log_median_income log_price log_cost_of_living
        log_mean_precip mean_summer_temp_cs mean_winter_temp_cs;
    bayes seed = 56549 ntu = 100 mintune = 20 maxtune = 20 nmc = 100000
        statistics = (summary interval) outpost = truckcount_post;
    prior intercept ~ normal(mean = 8.88, var = 10000);
    prior log_pop_bachelors log_pop_below_bachelors log_median_income
        log_cost_of_living log_mean_precip log_price ~ normal(mean = 0, var = 16);
    prior mean_summer_temp_cs mean_winter_temp_cs
        area_type_rural area_type_sub ~ normal(mean = 0, var = 7.62);
    prior log_price ~ normal(mean = -0.96, var = 0.25);
run;
```

Constructing the Predictive Distribution

For each draw from the posterior distribution:

1. Construct each new dealership's λ_i:
$$\log \lambda_i = x_i' \beta$$

2. Simulate each new dealership's predicted sales:
$$\text{sales}_i \sim \text{Poisson}(\lambda_i)$$

Before That: Rename Regression Coefficients

```
data truckcount_post;
    set truckcount_post;
    drop logpost loglike;
    rename log_pop_bachelors = b_log_pop_bachelors
        log_pop_below_bachelors = b_log_pop_below_bachelors
        log_median_income = b_log_median_income
        log_price = b_log_price
        log_cost_of_living = b_log_cost_of_living
        log_mean_precip = b_log_mean_precip
        mean_summer_temp_cs = b_mean_summer_temp_cs
        mean_winter_temp_cs = b_mean_winter_temp_cs;
run;
```

Simulate the Predictions

```
data truckcount_postpred;
    set truckcount_post;
    do j = 1 to nobs;
        set trucksales_missing point = j nobs = nobs;
        location = j;
        loglambda = intercept +
            log_pop_bachelors * b_log_pop_bachelors +
            log_pop_below_bachelors * b_log_pop_below_bachelors +
            log_median_income * b_log_median_income +
            log_price * b_log_price +
            log_cost_of_living * b_log_cost_of_living +
            log_mean_precip * b_log_mean_precip +
            mean_summer_temp_cs * b_mean_summer_temp_cs +
            mean_winter_temp_cs * b_mean_winter_temp_cs;
        if area_type = 'rural' then loglambda = loglambda + area_type_rural;
        if area_type = 'sub'   then loglambda = loglambda + area_type_sub;
        lambda = exp(loglambda);
        pred = rand('POISSON', lambda);
        output;
        end;
    keep iteration location pred;
run;
```

Reshape the Data Set

```
proc transpose data = truckcount_postpred
               out = truckcount_postpred
               prefix = pred;
    by iteration;
    id location;
    var pred;
run;

data truckcount_postpred;
    set truckcount_postpred;
    drop _NAME_;
run;
```

VIEWTABLE: Work.Truckcount_postpred

	Iteration	pred1	pred2	pred3	pred4	pred5
1	1	137	384	165	159	73
2	2	162	421	173	169	56
3	3	148	399	181	141	58
4	4	159	437	143	174	51
5	5	148	423	167	163	71

Summarize the Data Set

```
proc means data = truckcount_postpred maxdec = 2
           n mean std p5 p25 p50 p75 p95;
   var pred1-pred5;
run;
```

Variable	N	Mean	Std Dev	5th Pctl	25th Pctl	50th Pctl	75th Pctl	95th Pctl
pred1	100000	145.75	12.39	125.00	137.00	146.00	154.00	166.00
pred2	100000	418.90	22.03	383.00	404.00	419.00	434.00	455.00
pred3	100000	162.96	13.32	141.00	154.00	163.00	172.00	185.00
pred4	100000	172.71	14.03	150.00	163.00	173.00	182.00	196.00
pred5	100000	64.19	8.32	51.00	58.00	64.00	70.00	78.00

Use AUTOCALL Macros to Summarize

Requires SAS/STAT® Software

```
%postsum(data = truckcount_postpred, var = pred1-pred5)
%postint(data = truckcount_postpred, var = pred1-pred5)

%ess(data = truckcount_postpred, var = pred1-pred5)
%geweke(data = truckcount_postpred, var = pred1-pred5)
%heidel(data = truckcount_postpred, var = pred1-pred5)
%mcse(data = truckcount_postpred, var = pred1-pred5)
%raftery(data = truckcount_postpred, var = pred1-pred5)
```

Interval Statistics

Parameter	Alpha	CredibleLower	CredibleUpper	HPDLower	HPDUpper
pred1	0.05	121	170	121	169
pred2	0.05	376	462	374	460
pred3	0.05	137	189	135	187
pred4	0.05	145	201	143	198
pred5	0.05	49	81	47	79

Summary Statistics

Parameter	N	Mean	StdDev	P25	P50	P75
pred1	100000	145.749	12.3862	137	146	154
pred2	100000	418.905	22.0348	404	419	434
pred3	100000	162.959	13.3205	154	163	172
pred4	100000	172.710	14.0326	163	173	182
pred5	100000	64.189	8.3193	58	64	70

Thank you!

Contact Information
Matt.Simpson@sas.com

Code available on Github:

https://github.com/sascommunities/sas-global-forum-2020/tree/master/demos/SD313-Simpson-PostProcess

SAS® GLOBAL FORUM 2020