SAS® GLOBAL FORUM 2020

MARCH 29 – APRIL 1

USERS PROGRAM

Incorporating Auxiliary Information into Your Model Using Bayesian Methods in SAS® Econometrics

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

How many times have you fit a model and checked to see if the parameter estimates made sense?

You know something that your analysis does not take into account – why not improve it?

Bayesian methods enable you to take into account additional information through the prior

But doing this is not easy

Solution: Think real hard

The Game Plan

1. The Bayesian story
2. How to think about the prior
3. How to select the prior

See the paper for more detail and examples

The Bayesian Story

…and Its Discontents

Uncertainty is probability

- Probability as degree of belief

Consistent inferential framework

- Update beliefs with Bayes' rule

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

Where does the prior come from? Why does it seem made-up?

The Glib Bayesian Response

The prior comes from the same place as the likelihood

Darth_Vader.jpg

REDACTED

They're both made-up

Search your feelings. You know it to be true.

How to Think about the Prior

Together, the prior and likelihood are a model of your uncertainty about the problem

The big tricks:

- Focus on the distribution of observables
- Transform quantities in the model to make parameters easier to think about

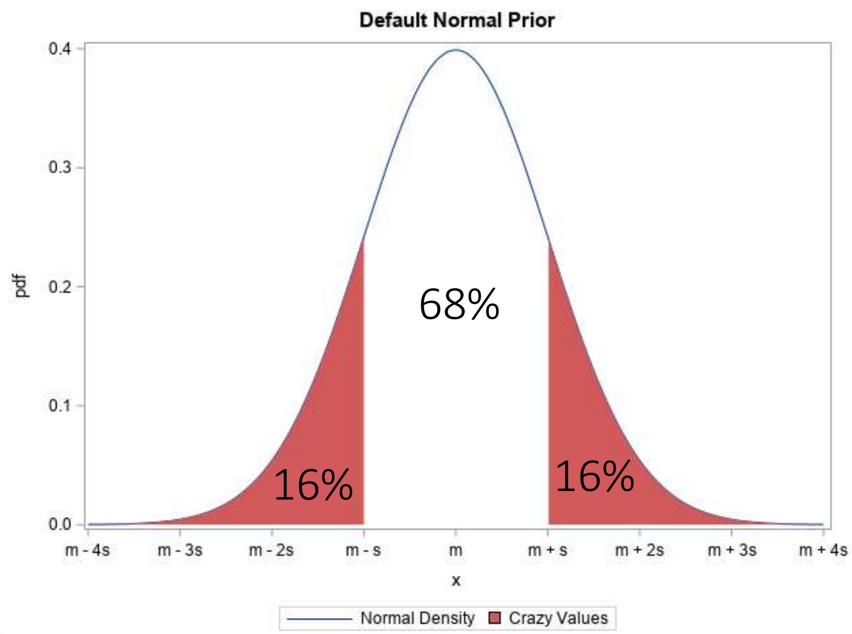
Start with a Reasonable Default Prior

Default priors:

- Weakly informative *for questions you care about*
- Spread out, but not too much

Starting point: $\theta \sim \mathrm{N}(m, s^2)$

- $m =$ the value you expect, or H_0
- $m \pm s =$ the most extreme value you think realistically possible



Example: 4×4 Truck Sales

Network of 100 dealerships

Want to predict a new dealership's sales using:

- Price
- Climate variables
- Demographic variables

Regression: Focus on price

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

Trick 1: Take Logs

If the response and covariate are both logged, the regression coefficient is an elasticity

A 1% change in x is associated with a
β% change in $\mathbf{sales}$

Null hypothesis / expected value? $\beta = 0$

Most extreme possible value? $\beta = \pm 4$ (room for disagreement)

So: $m = 0$ and $s = 4 - m = 4$ $\Longrightarrow$ $\beta \sim \mathrm{N}(0, 4^2)$

Trick 2: Standardize

$$\tilde{x}_{ij} = \frac{x_{ij} - \mathrm{MEAN}(\boldsymbol{x}_j)}{\mathrm{SD}(\boldsymbol{x}_j)}$$

A one-standard-deviation change in x is associated with a β change in $\log(\text{sales})$

Null hypothesis / expected value? $\beta = 0$

Most extreme possible value? $\beta = \pm 4 \times \mathrm{SD}[\log(\text{sales})]$

$m = 0$ and $s = 4 \times 0.69 - m = 2.76 \implies \beta \sim \mathrm{N}(0, 2.76^2)$

Trick 3: Base Cases

For classification variables, or in nonlinear and other complicated models, start with an intuitive base case

A change from the base group to a different group is associated with a β change in $\log(\text{sales})$.

Default choice: Same as a one-standard-deviation change in a continuous covariate

$\Longrightarrow$ Same prior: $\beta \sim \mathrm{N}(0, 2.76^2)$

Trick 4: Intercepts

The interpretation of the intercept depends on how other covariates were constructed and transformed

Default choice: Center on the classical intercept estimate

Prior SD set much larger than prior slope SDs

$$\Longrightarrow \qquad \beta \sim \mathrm{N}(8.88, 100^2)$$

Trick 5: Use Standard Deviations

Variances Are Bad

Variance

- Units are squared and dull

Standard Deviation

- Same cool units as the response

Amend the default prior to be truncated-normal: $\mathrm{N}^{+}(m, s^2)$

Default choice: $m = 0, \;\; s = \mathrm{SD}[\log(\text{sales})] = 0.69$

$$\Longrightarrow \qquad \sigma \sim \mathrm{N}^{+}(0, 0.69^2)$$

Informative Priors: What about $\beta_{\text{price}} > 0$

Vizzini.jpg REDACTED

Inigo.jpg REDACTED

Better choices:

$$\beta_{\text{price}} \sim \mathrm{N}(-1, 0.5^2)$$

$$\beta_{\text{price}} \sim \mathrm{N}(-2, 1^2)$$

Inconceivable!

You keep using that word

So force it to be negative?

I do not think it means what you think it means

Summary

How to think about the prior:

- Focus on distribution of observables
- Use transformations to make it easier
- These are tricks, not theorems!

Default prior: $\theta \sim \mathrm{N}(m, s^2)$

- Set $m = H_0$ or what you expect
- Set $s =$ the most extreme difference from m you think is possible
- Try alternative priors!

Fit the Model in PROC QLIM

```
proc qlim data = trucksales_transformed plots = none;
   class area_type;
   model log_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 = 72834 ntu = 100 mintune = 20 maxtune = 20 nmc = 10000;
   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 _sigma ~ normal(mean = 0, var = 0.48);
run; quit;
```

Fit the Model in PROC CQLIM

Coming Soon!

```
proc cqlim data = mycas.trucksales_transformed;
   class area_type;
   model log_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 = 72834 nsample = 10000
      sampler = rwm(ntu = 100 mintune = 20 maxtune = 20);
   prior intercept ~ normal(mean = 8.88, sd = 100);
   prior log_pop_bachelors log_pop_below_bachelors log_median_income
      log_cost_of_living log_mean_precip log_price ~ normal(mean = 0, sd = 4);
   prior mean_summer_temp_cs mean_winter_temp_cs
      area_type_rural area_type_sub ~ normal(mean = 0, sd = 2.72);
   prior _sigma ~ normal(mean = 0, sd = 0.69, lower = 0);
run; quit;
```

Fit the Model in the QLIM Action with Python or R

Coming Soon!

```
r = conn.qlim(
    table = 'trucksales_transformed',
    class_ = 'area_type',
    model = {'depVars' : 'log_sales',
             'effects' : ['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 = {'nsample' : 10000, 'seed' : 7284, 'priorsum' : True,
             'sampler' : {'method' : 'rwm',
                          'rwmOptions' : {'ntune' : 100, 'mintune' : 20, 'maxtune' : 20}}},
    prior = [{'parname' : 'Intercept',               'dist' : {'type' : 'normal', 'mean' : 8.88, 'sd' : 100}},
             {'parname' : 'area_type_rural',         'dist' : {'type' : 'normal', 'mean' : 0,    'sd' : 2.72}},
             {'parname' : 'area_type_sub',           'dist' : {'type' : 'normal', 'mean' : 0,    'sd' : 2.72}},
             {'parname' : 'mean_summer_temp_cs',     'dist' : {'type' : 'normal', 'mean' : 0,    'sd' : 2.72}},
             {'parname' : 'mean_winter_temp_cs',     'dist' : {'type' : 'normal', 'mean' : 0,    'sd' : 2.72}},
             {'parname' : 'log_pop_bachelors',       'dist' : {'type' : 'normal', 'mean' : 0,    'sd' : 4}},
             {'parname' : 'log_pop_below_bachelors', 'dist' : {'type' : 'normal', 'mean' : 0,    'sd' : 4}},
             {'parname' : 'log_median_income',       'dist' : {'type' : 'normal', 'mean' : 0,    'sd' : 4}},
             {'parname' : 'log_cost_of_living',      'dist' : {'type' : 'normal', 'mean' : 0,    'sd' : 4}},
             {'parname' : 'log_mean_precip',         'dist' : {'type' : 'normal', 'mean' : 0,    'sd' : 4}},
             {'parname' : 'log_price',               'dist' : {'type' : 'normal', 'mean' : 0,    'sd' : 4}},
             {'parname' : '_sigma',     'dist' : {'type' : 'normal', 'mean' : 0, 'sd' : 0.69, 'lower' : 0}}])
```

What Else Am I Doing at SAS Global Forum?

- "Incorporating Auxiliary Information into Your Model Using Bayesian Methods *in SAS Econometrics*"
  - Automatic implementation in SAS®
  - Super Demo

- "From Posterior to Postprocessing"
  - Posterior predictive inference – now let's forecast sales
  - Super Demo

Thank you!

Contact Information
Matt.Simpson@sas.com

Paper available on Github:

https://github.com/sascommunities/sas-global-forum-2020/tree/master/papers/4311-2020-Simpson

SAS® GLOBAL FORUM 2020