# Bayesian model averaging

Meghan K. Cain, Ph.D.
StataCorp, LLC

STATA 18

# Outline

- Motivating example
- Bayesian model selection and inference
- Bayesian model averaging
  - Theoretical introduction
  - Using `bmaregress`
  - Interpretation
  - Prior selection and sensitivity
    - On the model space
    - On the regression coefficients
  - Prediction

# Motivating example

```
. use ameshouses
(Ames house data)

. codebook, compact

Variable        Obs Unique      Mean     Min      Max  Label
----------------------------------------------------------------------------------------------------
yrsold         1451      5  2007.811    2006     2010  Year sold
soldage        1451    121  36.34045       0      136  Year sold - year built
soldremodel    1451     61  22.94418       0       60  Year sold - year of last remodel
lotfrontage    1193    109  69.98575      21      313  Linear feet of street connected to property
lotarea        1451   1070     10519    1300   215245  Lot size in square feet
overallqual    1451     10  6.099931       1       10  Rates the overall material and finish of the ...
overallcond    1451      9  5.569263       1        9  Rates the overall condition of the house
exterqual      1451      4  3.541006       1        4  Quality of exterior material
extercond      1451      5  4.735355       1        5  Present condition of exterior material
bsmtqual       1451      5   3.72295       1        5  Evaluates the height of the basement
kitchenqual    1451      4  3.339076       1        4  Kitchen quality
masvnrarea     1443    327  104.3319       0     1600  Masonry veneer area in square feet
bsmtfinsf1     1451    635  444.5651       0     5644  Basement Type 1 finished square feet
bsmtfinsf2     1451    144  46.83804       0     1474  Basement Type 2 finished square feet
bsmtunfsf      1451    777  567.1123       0     2336  Unfinished square feet of basement area
```

# Bayesian linear regression

```
. bayes, rseed(5924) burnin(10000): regress saleprice soldage

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  saleprice ~ regress(xb_saleprice,{sigma2})

Priors:
  {saleprice:soldage _cons} ~ normal(0,10000)                             (1)
                   {sigma2} ~ igamma(.01,.01)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_saleprice.

Bayesian linear regression                         MCMC iterations  =     20,000
Random-walk Metropolis-Hastings sampling           Burn-in          =     10,000
                                                   MCMC sample size =     10,000
                                                   Number of obs    =      1,451
                                                   Acceptance rate  =      .3535
                                                   Efficiency:  min =      .07257
```

# Bayesian linear regression

```
. bayes, rseed(5924) burnin(10000): regress saleprice soldage

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  saleprice ~ regress(xb_saleprice,{sigma2})

Priors:
  {saleprice:soldage _cons} ~ normal(0,10000)                              (1)
                   {sigma2} ~ igamma(.01,.01)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_saleprice.

Bayesian linear regression                       MCMC iterations  =     20,000
Random-walk Metropolis-Hastings sampling         Burn-in          =     10,000
                                                 MCMC sample size =     10,000
                                                 Number of obs    =      1,451
                                                 Acceptance rate  =      .3535
                                                 Efficiency:  min =     .07257
```

- Fitting model,

$$\texttt{saleprice} = \alpha + \beta\texttt{soldage} + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$

# Bayesian linear regression

```
. bayes, rseed(5924) burnin(10000): regress saleprice soldage

Burn-in ...
Simulation ...

Model summary
------------------------------------------------------------------------------
Likelihood:
  saleprice ~ regress(xb_saleprice,{sigma2})

Priors:
  {saleprice:soldage _cons} ~ normal(0,10000)                              (1)
                   {sigma2} ~ igamma(.01,.01)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_saleprice.

Bayesian linear regression                       MCMC iterations  =     20,000
Random-walk Metropolis-Hastings sampling         Burn-in          =     10,000
                                                 MCMC sample size =     10,000
                                                 Number of obs    =      1,451
                                                 Acceptance rate  =      .3535
                                                 Efficiency:  min =     .07257
```

- Fitting model,

$$\texttt{saleprice} = \alpha + \beta\texttt{soldage} + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$

- With priors,

$$\Pr(\alpha) \sim N(0{,}10000)$$
$$\Pr(\beta) \sim N(0{,}10000)$$

$$\Pr(\sigma^2) \sim IG(0.01{,}0.01)$$

```
Likelihood:
  saleprice ~ regress(xb_saleprice,{sigma2})

Priors:
  {saleprice:soldage _cons} ~ normal(0,10000)                          (1)
                   {sigma2} ~ igamma(.01,.01)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_saleprice.

Bayesian linear regression                       MCMC iterations  =     20,000
Random-walk Metropolis-Hastings sampling         Burn-in          =     10,000
                                                 MCMC sample size =     10,000
                                                 Number of obs    =      1,451
                                                 Acceptance rate  =      .3535
                                                 Efficiency:  min =     .07257
                                                              avg =      .1133
Log marginal-likelihood = -19628.377                          max =      .1763
```

| | Mean | Std. dev. | MCSE | Median | Equal-tailed [95% cred. | interval] |
|---|---|---|---|---|---|---|
| **saleprice** | | | | | | |
| soldage | 1252.152 | 74.73676 | 2.47559 | 1252.08 | 1107.63 | 1399.285 |
| _cons | 67.22493 | 99.29097 | 3.68577 | 66.78466 | -125.5595 | 262.4154 |
| sigma2 | 2.93e+10 | 1.14e+09 | 2.7e+07 | 2.92e+10 | 2.71e+10 | 3.16e+10 |

Note: Default priors are used for model parameters.

- Resulting in posteriors,

$$\Pr(\beta|\boldsymbol{y}) \propto L(\boldsymbol{\theta},\boldsymbol{y})\times\Pr(\beta)$$
$$\Pr(\alpha|\boldsymbol{y}) \propto L(\boldsymbol{\theta},\boldsymbol{y})\times\Pr(\alpha)$$
$$\Pr(\sigma^2|\boldsymbol{y}) \propto L(\boldsymbol{\theta},\boldsymbol{y})\times\Pr(\sigma^2)$$

# Bayesian model selection

```
. bayes, saving(m1, replace) rseed(5924) burnin(20000) mcmcsize(40000): regress saleprice soldage ///
>         overallcond overallqual

. estimates store m1


. bayes, saving(m2, replace) rseed(5924) burnin(20000) mcmcsize(40000): regress saleprice soldage ///
>         overallcond overallqual lotarea grlivarea bedroomabvgr

 . estimates store m2


. bayes, saving(m3, replace) rseed(5924) burnin(20000) mcmcsize(40000): regress saleprice soldage ///
>                 overallcond overallqual lotarea grlivarea bedroomabvgr i.fireplaces i.centralair kitchenqual
 . estimates store m3
```

# Bayesian model selection

```
. bayestest model m*, marglm(hmean)

Bayesian model tests
```

| | log(ML) | P(M) | P(M\|y) |
|---|---|---|---|
| m1 | **-1.95e+04** | **0.3333** | **0.0000** |
| m2 | **-1.78e+04** | **0.3333** | **1.0000** |
| m3 | **-1.78e+04** | **0.3333** | **0.0000** |

```
Note: Marginal likelihood (ML) is computed using
      harmonic-mean approximation.
```

# Bayesian model selection

```
. bayestest model m*, marglm(hmean)

Bayesian model tests
```

| | log(ML) | P(M) | P(M\|y) |
|---|---|---|---|
| m1 | **-1.95e+04** | **0.3333** | **0.0000** |
| m2 | **-1.78e+04** | **0.3333** | **1.0000** |
| m3 | **-1.78e+04** | **0.3333** | **0.0000** |

Note: Marginal likelihood (ML) is computed using
harmonic-mean approximation.

$$p(M|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|M) \times p(M)}{p(\boldsymbol{y})}$$

# Bayesian model inference

```
. estimates restore m2
(results m2 are active now)

. bayesstats summary, hpd

Posterior summary statistics                      MCMC sample size =     40,000
```

| | | | | | HPD | |
|---|---|---|---|---|---|---|
| | Mean | Std. dev. | MCSE | Median | [95% cred. | interval] |
| **saleprice** | | | | | | |
| soldage | -618.9771 | 36.90224 | 1.36451 | -619.7936 | -690.6499 | -546.7511 |
| overallcond | 191.7043 | 99.70763 | 4.9762 | 190.5521 | 3.717867 | 384.9718 |
| overallqual | 271.1908 | 98.90307 | 4.01183 | 269.3705 | 73.69656 | 458.1478 |
| lotarea | .8973271 | .1342066 | .004151 | .8970989 | .630286 | 1.146997 |
| grlivarea | 123.0915 | 1.465848 | .056339 | 123.0815 | 120.059 | 125.7533 |
| bedroomabvgr | -16.90889 | 98.68272 | 2.93159 | -16.8688 | -199.3762 | 184.591 |
| _cons | 32.25165 | 98.96611 | 2.84762 | 33.0959 | -158.1541 | 224.0952 |
| sigma2 | 2.52e+09 | 9.55e+07 | 1.1e+06 | 2.52e+09 | 2.34e+09 | 2.71e+09 |

# Bayesian model averaging

- BMA uses Bayes theorem to account for model uncertainty by considering model, $M_j$, as a random variable.

$$\Pr(M_j|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|M_j)\Pr(M_j)}{\Pr(\boldsymbol{y})}$$

- Considers all the models in the user-defined model space. With $k$ predictors there are $2^k$ possible models.

- Rather than selecting a single model, BMA takes model uncertainty into account when making inferences and predictions.

# Linear regression BMA

- We combine the likelihoods from $j$ linear regression models

$$\boldsymbol{y} = \alpha \mathbf{1}_n + \boldsymbol{X_j}\boldsymbol{\beta_j} + \boldsymbol{\varepsilon_j}$$

with the following prior distributions for models $M_j$ and regression parameters $\boldsymbol{\beta_j}$, $\alpha$, and $\sigma$.

$$\begin{aligned} M_j &\sim \Pr(M_j) \\ \boldsymbol{\beta_j}|\alpha, \sigma, M_j &\sim N_{kj}\{0, \sigma^2 g(\boldsymbol{X_j'}\boldsymbol{X_j})^{-1}\} \\ \alpha|\sigma, M_j &\propto 1 \\ \sigma|M_j &\propto \sigma^{-1} \end{aligned}$$

# BMA with `bmaregress`

```
. bmaregress saleprice soldage overallcond overallqual lotarea grlivarea bedroomabvgr ///
>          i.fireplaces i.centralair kitchenqual

Enumerating models ...
Computing model probabilities ...

Bayesian model averaging                          No. of obs         =      1,451
Linear regression                                 No. of predictors  =         11
Model enumeration                                             Groups =         11
                                                              Always =          0
Priors:                                           No. of models      =      2,048
  Models: Beta-binomial(1, 1)                         For CPMP >= .9 =          3
   Cons.: Noninformative                          Mean model size    =     10.174
   Coef.: Zellner's g
       g: Benchmark, g = 1,451                    Shrinkage, g/(1+g) =     0.9993
  sigma2: Noninformative                          Mean sigma2        = 1.332e+09

------------------------------------------------------------------------------
   saleprice |      Mean   Std. dev.                         Group        PIP
-------------+----------------------------------------------------------------
     soldage | -583.2091    45.72816                             1          1
```

| saleprice | Mean | Std. dev. | Group | PIP |
|---|---|---|---|---|
| soldage | -583.2091 | 45.72816 | 1 | 1 |
| overallqual | 17819.47 | 1184.367 | 3 | 1 |
| grlivarea | 62.65485 | 3.145142 | 5 | 1 |
| kitchenqual | -16009.7 | 1435.584 | 11 | 1 |
| lotarea | .7820294 | .1027278 | 4 | 1 |
| overallcond | 5700.984 | 975.4184 | 2 | 1 |
| fireplaces | | | | |
| 2 | 24327.11 | 4274.342 | 8 | 1 |
| bedroomabvgr | -7469.643 | 1499.924 | 6 | .99997 |
| fireplaces | | | | |
| 3 | -69610.76 | 17107.83 | 9 | .99863 |
| 1 | 5797.346 | 2920.783 | 7 | .89739 |
| centralair | | | | |
| Yes | 1059.696 | 2952.486 | 10 | .27768 |
| **Always** | | | | |
| _cons | 28277.48 | 10570.24 | 0 | 1 |

Note: Coefficient posterior means and std. dev. estimated from 2,048 models.
Note: Default priors are used for models and parameter *g*.

| saleprice | Mean | Std. dev. | Group | PIP |
|---|---|---|---|---|
| soldage | -583.2091 | 45.72816 | 1 | 1 |
| overallqual | 17819.47 | 1184.367 | 3 | 1 |
| grlivarea | 62.65485 | 3.145142 | 5 | 1 |
| kitchenqual | -16009.7 | 1435.584 | 11 | 1 |
| lotarea | .7820294 | .1027278 | 4 | 1 |
| overallcond | 5700.984 | 975.4184 | 2 | 1 |
| fireplaces | | | | |
| 2 | 24327.11 | 4274.342 | 8 | 1 |
| bedroomabvgr | -7469.643 | 1499.924 | 6 | .99997 |
| fireplaces | | | | |
| 3 | -69610.76 | 17107.83 | 9 | .99863 |
| 1 | 5797.346 | 2920.783 | 7 | .89739 |
| centralair | | | | |
| Yes | 1059.696 | 2952.486 | 10 | .27768 |
| **Always** | | | | |
| _cons | 28277.48 | 10570.24 | 0 | 1 |

Note: Coefficient posterior means and std. dev. estimated from 2,048 models.
Note: Default priors are used for models and parameter *g*.

$$PIP(\boldsymbol{X_k}) = \sum_j I(\boldsymbol{X_k} \in M_j)\Pr(M_j|\boldsymbol{y})$$

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{BMA}} = \sum_j \widehat{\boldsymbol{\beta}}_{\boldsymbol{M_j}}\Pr(M_j|\boldsymbol{y})$$

# BMA with bmaregress

```
. bmaregress saleprice soldage-poolarea i.fireplaces i.seasonsold i.foundation    ///
>          i.centralair, rseed(92823)

Burn-in ...
Simulation ...
Computing model probabilities ...

Bayesian model averaging                        No. of obs         =      1,113
Linear regression                               No. of predictors  =         46
MC3 sampling                                                Groups =         46
                                                            Always =          0
                                                No. of models      =        417
                                                    For CPMP >= .9 =         78
Priors:                                         Mean model size    =     15.645
  Models: Beta-binomial(1, 1)                   Burn-in            =      2,500
   Cons.: Noninformative                        MCMC sample size   =     10,000
   Coef.: Zellner's g                           Acceptance rate    =     0.0729
       g: Benchmark, g = 2,116                  Shrinkage, g/(1+g) =     0.9995
  sigma2: Noninformative                        Mean sigma2        = 1.205e+09

Sampling correlation = 0.8517
```

| saleprice | Mean | Std. dev. | Group | PIP |
|---|---|---|---|---|
| soldage | -250.683 | 54.20822 | 1 | 1 |
| lotarea | .7073138 | .1435761 | 4 | 1 |
| overallqual | 13611.92 | 1433.411 | 5 | 1 |
| overallcond | 6991.137 | 1135.485 | 6 | 1 |
| kitchenqual | -11120.51 | 1736.539 | 10 | 1 |
| kitchenabvgr | -27820.47 | 6192.748 | 25 | 1 |
| garagecars | 17026.02 | 2351.316 | 29 | 1 |
| fireplaces | | | | |
| 3 | -121188.6 | 18575.55 | 39 | 1 |
| exterqual | -10110.12 | 2415.279 | 7 | .99976 |
| bsmtqual | -5104.976 | 1188.016 | 9 | .99945 |
| fireplaces | | | | |
| 2 | 21484.03 | 4791.198 | 38 | .99352 |
| masvnrarea | 23.29787 | 8.219115 | 11 | .95217 |
| bsmtfinsf1 | 16.6102 | 7.210094 | 12 | .86215 |
| grlivarea | 35.89608 | 18.95823 | 19 | .80467 |
| stflrsf | 14.2591 | 20.34168 | 16 | .4932 |
| ndflrsf | 5.478727 | 18.03836 | 17 | .38956 |
| totrmsabvgrd | 1303.966 | 1868.175 | 26 | .3678 |
| totalbsmtsf | 4.72351 | 9.635595 | 15 | .23224 |
| bsmtunfsf | -2.473556 | 6.365035 | 14 | .14614 |

# BMA with bmaregress

```
. bmaregress saleprice soldage-poolarea i.fireplaces i.seasonsold i.foundation    ///
>         i.centralair, groupfv rseed(92823)

Burn-in ...
Simulation ...
Computing model probabilities ...

Bayesian model averaging                        No. of obs         =      1,113
Linear regression                               No. of predictors  =         46
MC3 sampling                                                Groups =         40
                                                            Always =          0
                                                No. of models      =        342
                                                    For CPMP >= .9 =         73
Priors:                                         Mean model size    =     16.471
  Models: Beta-binomial(1, 1)                   Burn-in            =      2,500
   Cons.: Noninformative                        MCMC sample size   =     10,000
   Coef.: Zellner's g                           Acceptance rate    =     0.0772
       g: Benchmark, g = 2,116                  Shrinkage, g/(1+g) =     0.9995
  sigma2: Noninformative                        Mean sigma2        =  1.205e+09

Sampling correlation = 0.7747
```

| saleprice | Mean | Std. dev. | Group | PIP |
|---|---|---|---|---|
| soldage | -255.1568 | 54.6001 | 5 | 1 |
| overallqual | 13245.2 | 1460.802 | 9 | 1 |
| overallcond | 7028.775 | 1137.665 | 10 | 1 |
| kitchenqual | -11215.49 | 1736.509 | 14 | 1 |
| grlivarea | 43.74483 | 7.084451 | 23 | 1 |
| garagecars | 17005.77 | 2361.825 | 33 | 1 |
| | | | | |
| fireplaces | | | | |
| 1 | 3243.485 | 2540.853 | 1 | 1 |
| 2 | 23921.89 | 4768.517 | 1 | 1 |
| 3 | -116955.2 | 18675.76 | 1 | 1 |
| | | | | |
| lotarea | .7084268 | .1448688 | 8 | .99993 |
| exterqual | -10227.9 | 2432.821 | 11 | .99967 |
| kitchenabvgr | -25912.97 | 6278.038 | 29 | .99781 |
| bsmtqual | -5133.269 | 1211.252 | 13 | .9976 |
| masvnrarea | 23.21975 | 8.279709 | 15 | .94993 |
| bsmtfinsf1 | 15.02891 | 8.579624 | 16 | .77529 |
| totalbsmtsf | 7.623932 | 11.34056 | 19 | .37014 |
| totrmsabvgrd | 1133.925 | 1761.972 | 30 | .33231 |
| stflrsf | 3.746267 | 6.103564 | 20 | .30286 |
| bsmtunfsf | -4.01715 | 7.687266 | 18 | .23843 |
| ndflrsf | -2.228069 | 4.941738 | 21 | .18813 |
| screenporch | 3.191943 | 11.88422 | 39 | .082965 |
| wooddecksf | 1.056389 | 4.887271 | 35 | .055113 |

# Model and variable-inclusion summaries

```
. bmastats models

Computing model probabilities ...

Model summary                                     Number of models:
                                                           Visited = 342
                                                          Reported =   5
```

| | Analytical PMP | Frequency PMP | Model size |
|---|---|---|---|
| **Rank** | | | |
| 1 | .1209 | .0451 | 16 |
| 2 | .0849 | .0345 | 16 |
| 3 | .08232 | .1315 | 16 |
| 4 | .07331 | .0309 | 15 |
| 5 | .06329 | .0213 | 17 |

Note: Using analytical PMP for model ranking.

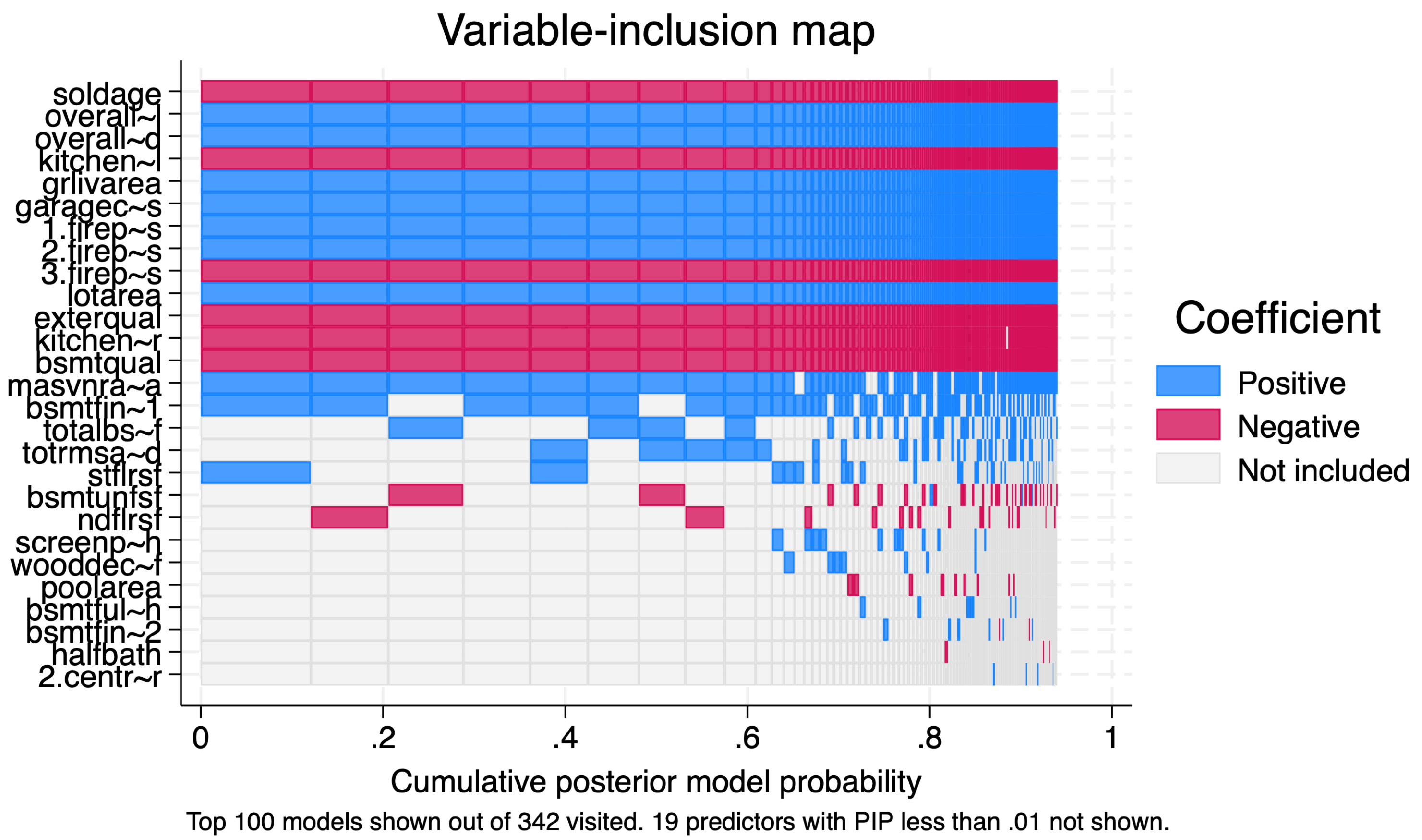
Variable-inclusion summary

Variable-inclusion summary

| | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|
| soldage | x | x | x | x | x |
| lotarea | x | x | x | x | x |
| overallqual | x | x | x | x | x |
| overallcond | x | x | x | x | x |
| exterqual | x | x | x | x | x |
| bsmtqual | x | x | x | x | x |
| kitchenqual | x | x | x | x | x |
| masvnrarea | x | x | x | x | x |
| bsmtfinsf1 | x | x | | x | x |
| stflrsf | x | | | | x |
| grlivarea | x | x | x | x | x |
| kitchenabvgr | x | x | x | x | x |
| garagecars | x | x | x | x | x |
| fireplaces | | | | | |
| 1 | x | x | x | x | x |
| 2 | x | x | x | x | x |
| 3 | x | x | x | x | x |
| ndflrsf | | x | | | |
| bsmtunfsf | | | x | | |
| totalbsmtsf | | | x | | |
| totrmsabvgrd | | | | | x |

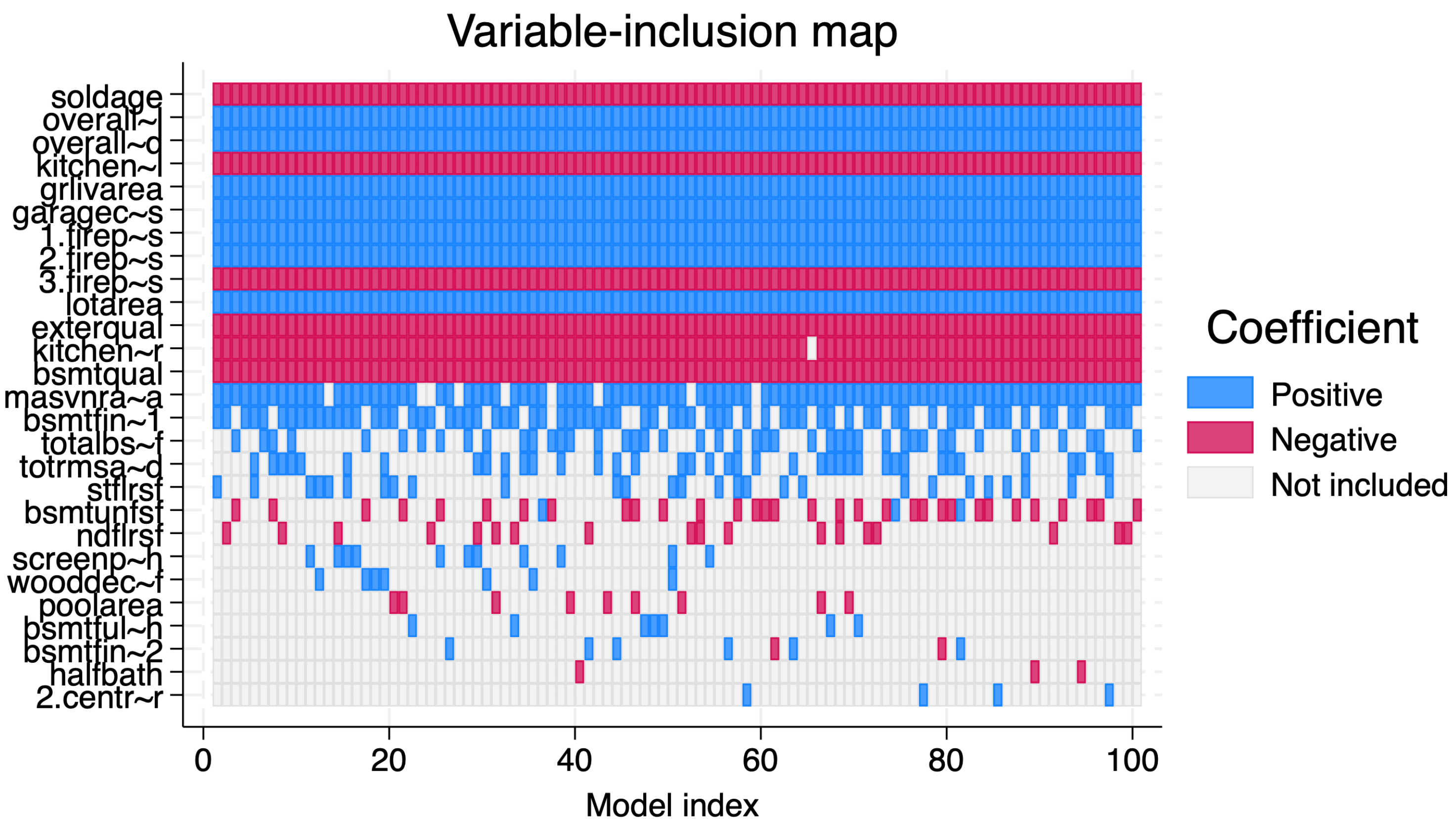
# Variable-inclusion map

```
. bmagraph varmap
```



# Variable-inclusion map

```
. bmagraph varmap, equalwidth
```



Top 100 models shown out of 342 visited. 19 predictors with PIP less than .01 not shown.

# Jointness

```
. bmastats jointness totalbsmtsf bsmtunfsf

Computing model probabilities ...

Variables: totalbsmtsf bsmtunfsf

─────────────────────────────────────────────
                               │  Jointness
───────────────────────────────┼─────────────
             Doppelhofer–Weeks │   4.639521
              Ley–Steel type 1 │   .6021957
              Ley–Steel type 2 │   1.513799
                      Yule's Q │   .9808603
─────────────────────────────────────────────
Notes: Using analytical PMPs. See
       thresholds.
```

# Jointness

```
. bmastats jointness totalbsmtsf bsmtunfsf

Computing model probabilities ...

Variables: totalbsmtsf bsmtunfsf

─────────────────────────────────────
                     │   Jointness
─────────────────────┼───────────────
     Doppelhofer–Weeks│   4.639521
      Ley–Steel type 1│   .6021957
      Ley–Steel type 2│   1.513799
              Yule's Q│   .9808603
─────────────────────────────────────
Notes: Using analytical PMPs. See
       thresholds.
```

| DW | Interpretation |
|---|---|
| $(-\infty, -2)$ | Strong disjointness |
| $[-2, -1)$ | Significant disjointness |
| $[-1, 1]$ | Independent inclusion |
| $(1, 2]$ | Significant jointness |
| $(2, \infty)$ | Strong jointness |

| LS2 | LS1* | Interpretation |
|---|---|---|
| $[0, 0.01)$ | $[0, 0.01)$ | Decisive disjointness |
| $[0.01, 0.03)$ | $[0.01, 0.03)$ | Very strong disjointness |
| $[0.03, 0.1)$ | $[0.03, 0.09)$ | Strong disjointness |
| $[0.1, 0.22)$ | $[0.09, 0.18)$ | Favorable disjointness |
| $[0.22, 3)$ | $[0.18, 0.75)$ | Independent inclusion |
| $[3, 10)$ | $[0.75, 0.91)$ | Favorable jointness |
| $[10, 30)$ | $[0.91, 0.97)$ | Strong jointness |
| $[30, 100)$ | $[0.97, 0.99)$ | Very strong jointness |
| $[100, \infty)$ | $[0.99, 1]$ | Decisive jointness |

# Prior selection

```
. bmaregress saleprice soldage-poolarea i.fireplaces i.seasonsold i.foundation    ///
>         i.centralair, groupfv rseed(92823)

Burn-in ...
Simulation ...
Computing model probabilities ...

Bayesian model averaging                       No. of obs         =      1,113
Linear regression                              No. of predictors  =         46
MC3 sampling                                               Groups =         40
                                                           Always =          0
                                               No. of models      =        342
                                                   For CPMP >= .9 =         73
Priors:                                        Mean model size    =     16.471
  Models: Beta-binomial(1, 1)                  Burn-in            =      2,500
   Cons.: Noninformative                       MCMC sample size   =     10,000
   Coef.: Zellner's g                          Acceptance rate    =     0.0772
       g: Benchmark, g = 2,116                 Shrinkage, g/(1+g) =     0.9995
  sigma2: Noninformative                       Mean sigma2        =  1.205e+09
```

# Prior selection

```
. bmaregress saleprice soldage-poolarea i.fireplaces i.seasonsold i.foundation    ///
>         i.centralair, groupfv rseed(92823)

Burn-in ...
Simulation ...
Computing model probabilities ...

Bayesian model averaging                       No. of obs         =      1,113
Linear regression                              No. of predictors  =         46
MC3 sampling                                               Groups =         40
                                                           Always =          0
                                               No. of models      =        342
                                                   For CPMP >= .9 =         73
Priors:                                        Mean model size    =     16.471
  Models: Beta-binomial(1, 1)                  Burn-in            =      2,500
   Cons.: Noninformative                       MCMC sample size   =     10,000
   Coef.: Zellner's g                          Acceptance rate    =     0.0772
       g: Benchmark, g = 2,116                 Shrinkage, g/(1+g) =     0.9995
  sigma2: Noninformative                       Mean sigma2        =  1.205e+09
```

# Priors

- Prior for model space

  `mprior()`

| | |
|---|---|
| `betabinomial` | model size |
| `uniform` | model space |
| `binomial` | inclusion probabilities |

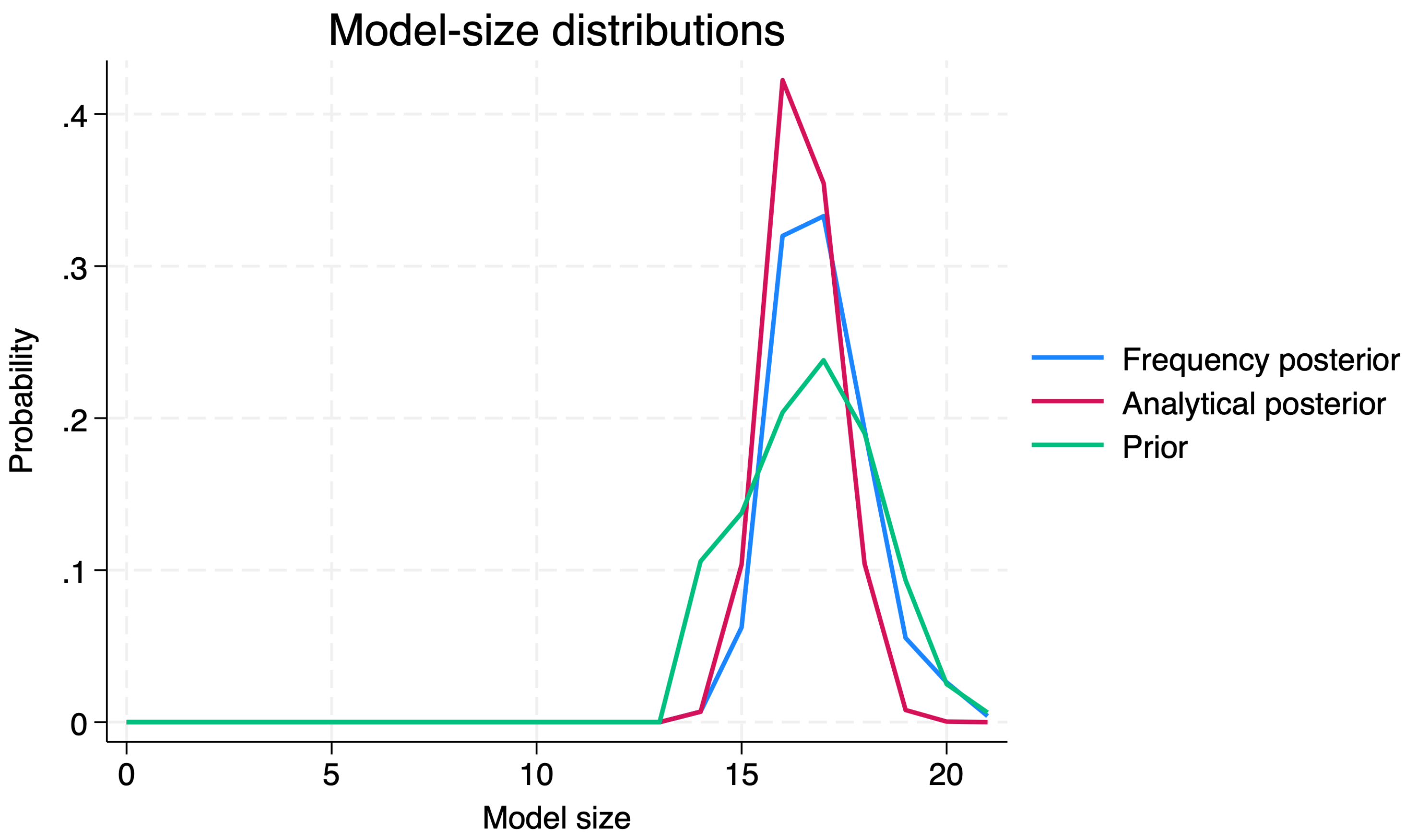
- Prior for Zellner's *g* (shrinkage of regression coefficients)

  `gprior()`

| | | | |
|---|---|---|---|
| `bench` (default) | `sqrtn` | `betashrink` *#1 #2* | `hypergn` *#* |
| `uip` | `fixed` *#* | `betabench` *#* | `zsiow` |
| `ric` | `ebl` | `hyperg` *#* | `robust` |

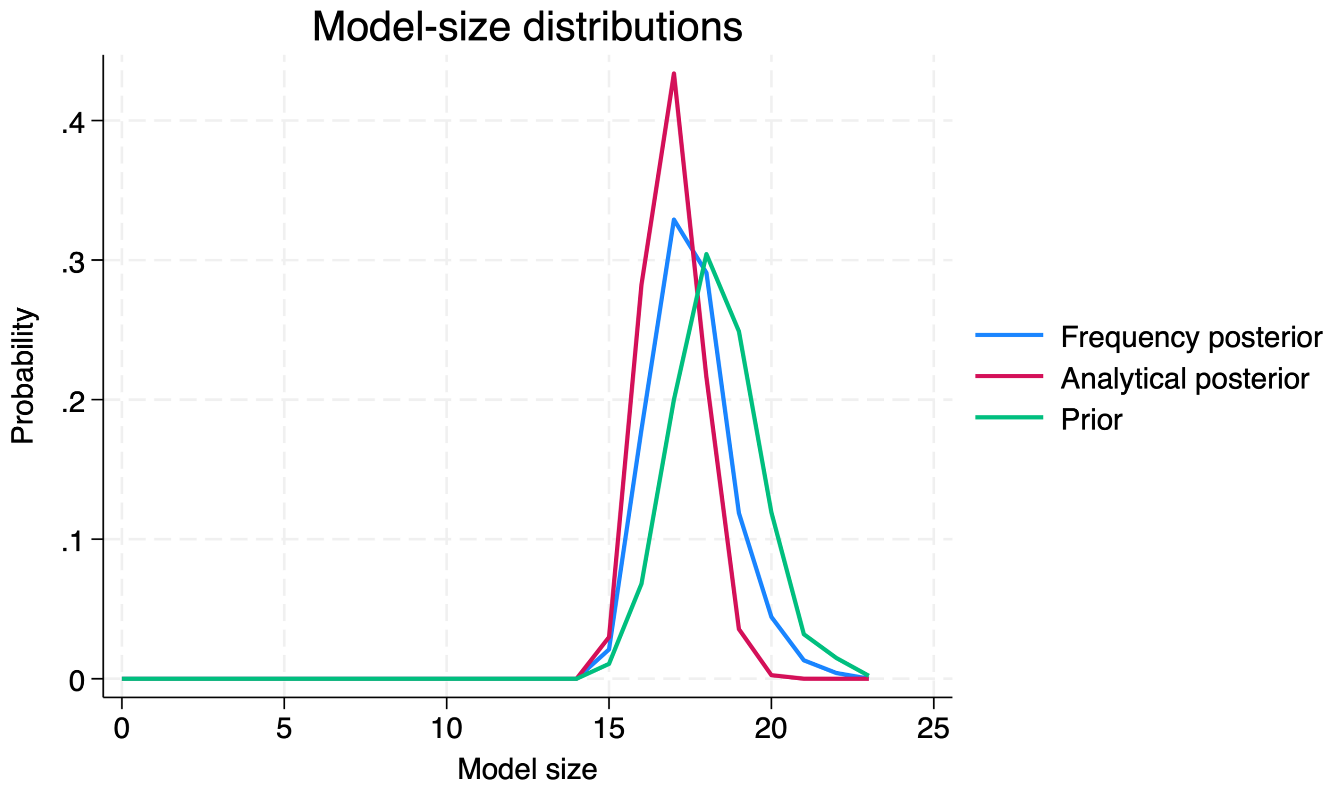
# Priors: Model space

```
. bmagraph msize
```



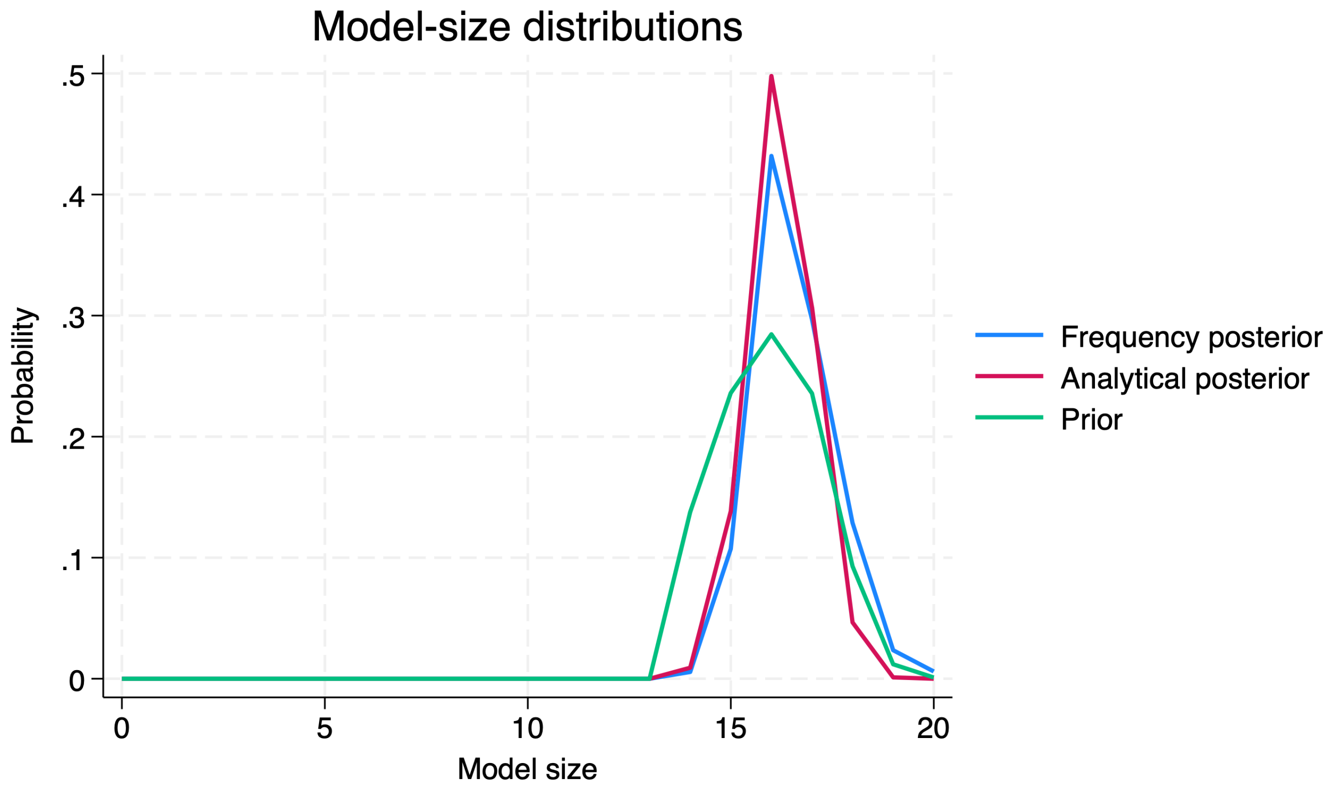
# Priors: Model space

```
. bmaregress saleprice soldage-poolarea i.fireplaces i.seasonsold i.foundation    ///
>         i.centralair, groupfv rseed(92823) mprior(uniform)
. bmagraph msize
```



# Priors: Model space

```
. bmaregress saleprice soldage-poolarea i.fireplaces i.seasonsold i.foundation    ///
>         i.centralair, groupfv rseed(92823) mprior(binomial .3)
. bmagraph msize
```



# Priors: Zellner's *g*

- Specified with option `gprior()`

- Zellner's *g* can be fixed

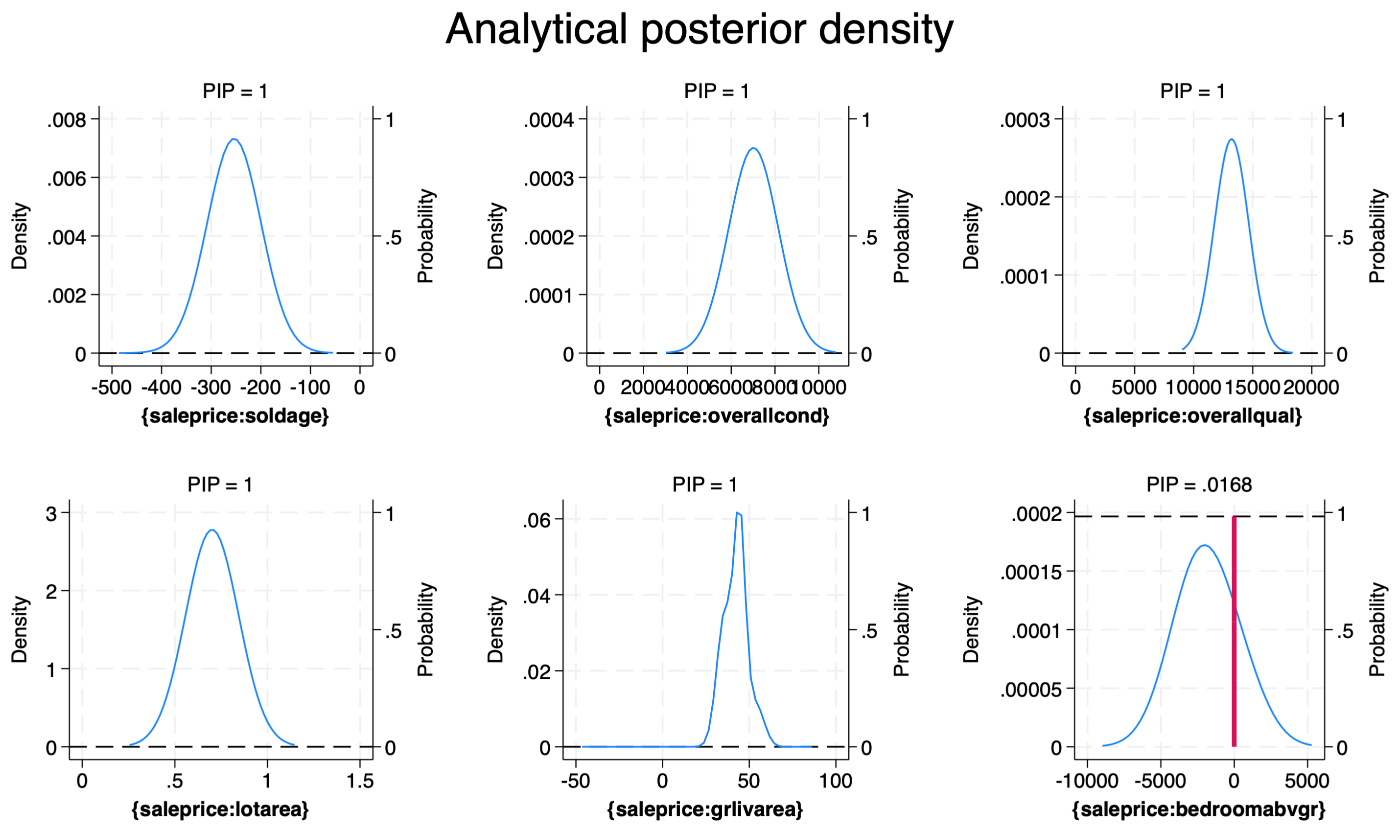
| | |
|---|---|
| `bench` (default) | `sqrtn` |
| `uip` | `fixed` # |
| `ric` | `ebl` |

or random

| | |
|---|---|
| `betashrink` *#1 #2* | `hypergn` # |
| `betabench` # | `zsiow` |
| `hyperg` # | `robust` |

# Priors: Zellner's *g*

```
. bmacoefsample, rseed(92823)
. bmagraph coefdensity soldage overallcond overallqual lotarea grlivarea bedroomabvgr, combine
```



# Priors: Zellner's *g*

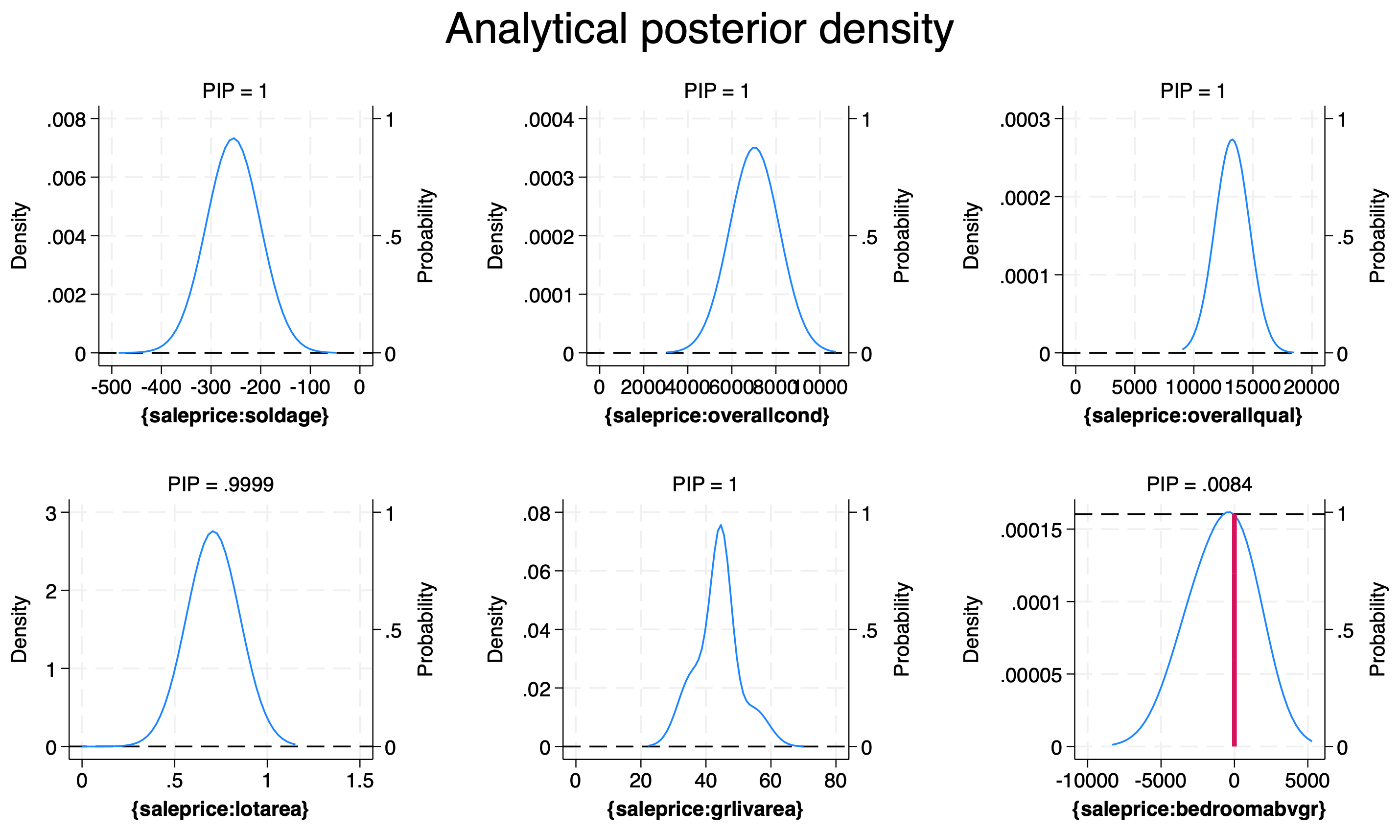
```
. bmaregress saleprice soldage-poolarea i.fireplaces i.seasonsold i.foundation i.centralair, ///
>          groupfv rseed(92823) gprior(ric) saving(bma_ric, replace)

Burn-in ...
Simulation ...
Computing model probabilities ...

Bayesian model averaging                        No. of obs         =     1,113
Linear regression                               No. of predictors  =        46
MC3 sampling                                                Groups =        40
                                                            Always =         0
                                                No. of models      =       342
                                                    For CPMP >= .9 =        73
Priors:                                         Mean model size    =    16.471
  Models: Beta-binomial(1, 1)                   Burn-in            =     2,500
   Cons.: Noninformative                        MCMC sample size   =    10,000
   Coef.: Zellner's g                           Acceptance rate    =    0.0772
       g: Risk inflation, g = 2,116             Shrinkage, g/(1+g) =    0.9995
  sigma2: Noninformative                        Mean sigma2        = 1.205e+09
```

# Priors: Zellner's *g*

```
. bmacoefsample, rseed(92823)
. bmagraph coefdensity soldage overallcond overallqual lotarea grlivarea bedroomabvgr, combine
```



# Priors: Zellner's *g*

```
. bmaregress saleprice soldage-poolarea i.fireplaces i.seasonsold i.foundation i.centralair, ///
>         groupfv rseed(92823) gprior(zsiow) saving(bma_zsiow, replace)

Burn-in ...
Simulation ...
Computing model probabilities ...

Bayesian model averaging                         No. of obs         =      1,113
Linear regression                                No. of predictors =         46
MC3 and adaptive MH sampling                                Groups =         40
                                                            Always =          0
                                                 No. of models      =        916
                                                    For CPMP >= .9 =        530
Priors:                                          Mean model size    =     18.534
  Models: Beta-binomial(1, 1)                    Burn-in            =      2,500
   Cons.: Noninformative                         MCMC sample size   =     10,000
   Coef.: Zellner's g                            Acceptance rate    =     0.5167
       g: Zellner-Siow
  sigma2: Noninformative                         Mean sigma2        = 1.213e+09
```

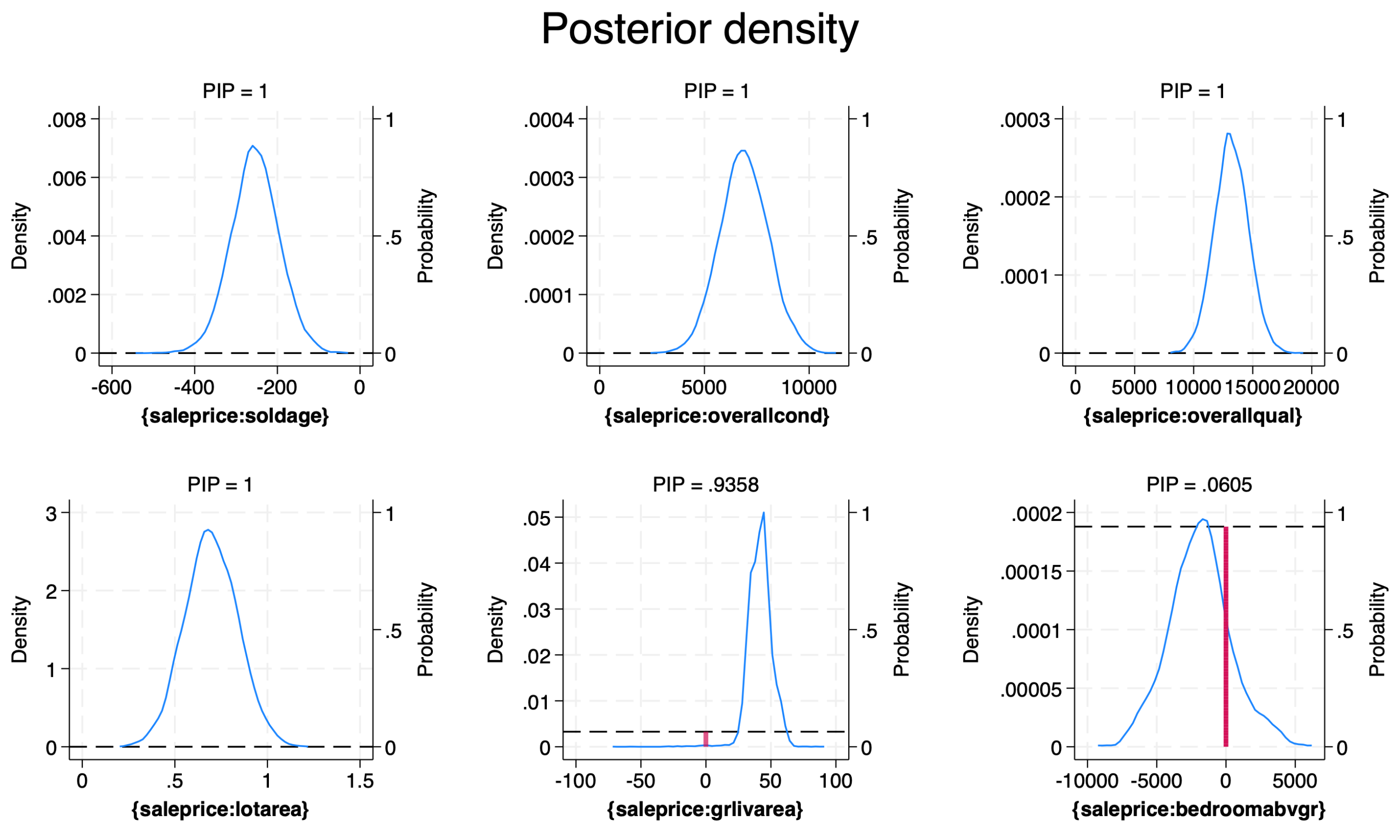
| | | | | |
|---|---|---|---|---|
| soldremodel | .8544056 | 19.80145 | 6 | .0582 |
| openporchsf | -.3192711 | 4.517522 | 36 | .0536 |
| garagearea | .1812007 | 2.647387 | 34 | .0531 |
| enclosedporch | -.342919 | 4.496452 | 37 | .0466 |
| fullbath | 62.32316 | 706.5014 | 26 | .046 |
| extercond | 39.92509 | 367.2668 | 12 | .0362 |
| lotfrontage | .5414932 | 10.70252 | 7 | .0361 |
| fireplaces | -1032.129 | 6291.884 | 31 | .0269 |
| seasonsold | | | | |
| Summer | 43.01961 | 484.6776 | 2 | .0111 |
| Fall | -40.45628 | 512.7442 | 2 | .0111 |
| Winter | -15.64621 | 414.9722 | 2 | .0111 |
| **Always** | | | | |
| _cons | 64993.44 | 58821.62 | 0 | 1 |

Note: Coefficient posterior means and std. dev. estimated from 916 models.
Note: Default prior is used for models.
Note: 3 predictors with PIP less than .01 not shown.

| | Mean | Std. dev. | MCSE | Median | Equal-tailed [95% cred. | interval] |
|---|---|---|---|---|---|---|
| g | 362.9077 | 134.1596 | 3.79033 | 336.5314 | 185.8244 | 697.7702 |
| Shrinkage | .9969267 | .0009978 | .000029 | .9970373 | .9946474 | .9985689 |

# Priors: Zellner's *g*

```
. bmacoefsample, rseed(92823)
. bmagraph coefdensity soldage overallcond overallqual lotarea grlivarea bedroomabvgr, combine
```



# Prediction

- Model averaging can be viewed as an ensemble method that improves predictive performance using optimal combinations in the space of considered candidate models.

- First, refit the model to sales prior to 2010 and store results,

```
. bmaregress saleprice soldage-poolarea i.fireplaces i.seasonsold i.foundation i.centralair       ///
>         if yrsold<2010, rseed(5924) saving(bma, replace)
. estimates store bma

. bmacoefsample, rseed(5924) saving(bmacoefs, replace)
```

- Then compute posterior predictive means of sales price

```
. bmapredict pmean_bma, mean
note: computing analytical posterior predictive means.
```

# Prediction

- Let's compare these predictions to our original linear regression model we chose:

```
. bmaregress saleprice (soldage overallcond overallqual lotarea grlivarea bedroomabvgr, always) ///
>         if yrsold<2010, saving(regress, replace)
. estimates store regress


. bmacoefsample, rseed(5924) saving(regcoefs, replace)

. bmapredict pmean_reg, mean
note: computing analytical posterior predictive means.
```

# Prediction: Mean squared error

```
. generate sqerr_bma = (saleprice-pmean_bma)^2 if yrsold==2010

. generate sqerr_reg = (saleprice-pmean_reg)^2 if yrsold==2010


. summarize sq* if !missing(sqerr_bma)
```

| Variable | Obs | Mean | Std. dev. | Min | Max |
|---|---|---|---|---|---|
| sqerr_bma | 129 | 9.54e+08 | 3.46e+09 | 1814.228 | 3.58e+10 |
| sqerr_reg | 129 | 1.75e+09 | 7.26e+09 | 80288.11 | 7.75e+10 |

# Prediction: Log predictive scores

$$LPS_{y*} = -log\left\{\sum_{j} \Pr\left(M_j \middle| \boldsymbol{y}\right) f(\boldsymbol{y}^* | M_j, \boldsymbol{y})\right\}$$

where $f(y^*|M_j, \boldsymbol{y})$ is the conditional posterior predictive distribution

```
. bmastats lps bma regress if yrsold==2010

Log predictive-score (LPS) for bma

Number of observations =        172
Posterior mean sigma2  =   1.25e+09
Entropy                =   11.89322
```

| | Mean | Minimum | Maximum |
|---|---|---|---|
| LPS | 11.77221 | 11.39516 | 24.80956 |

```
Note: Using analytical PMPs.

Log predictive-score (LPS) for regress

Number of observations =        172
Posterior mean sigma2  =   1.50e+09
Entropy                =   11.98334
```

| | Mean | Minimum | Maximum |
|---|---|---|---|
| LPS | 11.97647 | 11.48399 | 36.67776 |

```
Note: Using analytical PMPs.
```

STATA® 18