Bayesian variable selection

Dr. Jarad Niemi

Iowa State University

September 9, 2021

Bayesian regression

Consider the model

$$y = X\beta + \epsilon$$

with

$$\epsilon \sim N(0, \sigma^2 \mathrm{I})$$

where

- y is a vector of length n
- β is an unknown vector of length p
- X is a known $n \times p$ design matrix
- σ^2 is an unknown scalar

For a given design matrix X, we are interested in the posterior

$$p(\beta, \sigma^2 | y),$$

but we may also be interested in which columns of X should be included, i.e. what explanatory variables should we keep in the model.

Default Bayesian regression

Assume the standard noninformative prior

$$p(\beta, \sigma^2) \propto 1/\sigma^2$$

then the posterior is

$$\begin{aligned}
p(\beta, \sigma^2|y) &= p(\beta|\sigma^2, y)p(\sigma^2|y) \\
\beta|\sigma^2, y &\sim N(\hat{\beta}_{MLE}, \sigma^2 V_\beta) \\
\sigma^2|y &\sim IG\left(\tfrac{n-p}{2}, \tfrac{[n-p]s^2}{2}\right) \\
\beta|y &\sim t_{n-p}(\hat{\beta}_{MLE}, s^2 V_\beta)
\end{aligned}$$

$$\begin{aligned}
V_\beta &= (X^\top X)^{-1} \\
\hat{\beta}_{MLE} &= V_\beta X^\top y \\
s^2 &= \tfrac{1}{n-p}(y - X\hat{\beta}_{MLE})^\top (y - X\hat{\beta}_{MLE})
\end{aligned}$$

The posterior is proper if $n > p$ and $\text{rank}(X) = p$.

Information about chirps per 15 seconds

Let

- Y_i is the average number of chirps per 15 seconds and
- X_i is the temperature in Fahrenheit.

And we assume

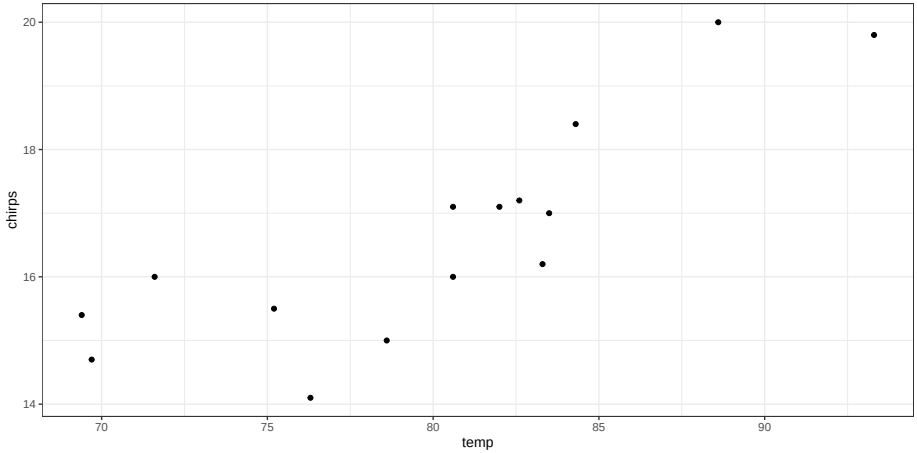
$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

then

- β_0 is the expected number of chirps at 0 degrees Fahrenheit
- β_1 is the expected increase in number of chirps (per 15 seconds) for each degree increase in Fahrenheit.

Cricket chirps

As an example, consider the relationship between the number of cricket chirps (in 15 seconds) and temperature (in Fahrenheit). From example in `LearnBayes::blinreg`.



Default Bayesian regression

```
summary(m <- lm(chirps~temp))


##
## Call:
## lm(formula = chirps ~ temp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74107 -0.58123  0.02956  0.58250  1.50608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.61521    3.14434  -0.196 0.847903
## temp         0.21568    0.03919   5.504 0.000102 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9849 on 13 degrees of freedom
## Multiple R-squared:  0.6997,Adjusted R-squared:  0.6766
## F-statistic: 30.29 on 1 and 13 DF,  p-value: 0.0001015


confint(m) # Credible intervals


##                 2.5 %    97.5 %
## (Intercept) -7.4081577 6.1777286
## temp         0.1310169 0.3003406
```

Fully conjugate subjective Bayesian inference

If we assume the following normal-inverse-gamma prior,

$$\beta|\sigma^2 \sim N(b_0, \sigma^2 B_0) \qquad \sigma^2 \sim IG(a, b)$$

then the posterior is

$$\beta|\sigma^2, y \sim N(b_n, \sigma^2 B_n) \qquad \sigma^2|y \sim IG(a', b')$$

with

$$\begin{aligned} B_n^{-1} &= B_0^{-1} + \tfrac{1}{\sigma^2} X^\top X \\ b_n &= B_n^{-1}\left[B_0^{-1} b_0 + \tfrac{1}{\sigma^2} X^\top y\right] \\ a' &= a + \tfrac{n}{2} \\ b' &= b + \tfrac{1}{2}(y - Xb_0)^\top (XB_0X^\top + \mathrm{I})^{-1}(y - Xb_0) \end{aligned}$$

Information about chirps per 15 seconds

Let

- Y_i is the average number of chirps per 15 seconds and
- X_i is the temperature in Fahrenheit.

And we assume

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

then

- β_0 is the expected number of chirps at 0 degrees Fahrenheit
- β_1 is the expected increase in number of chirps (per 15 seconds) for each degree increase in Fahrenheit.

Perhaps a reasonable prior is $p(\beta_0, \beta_1, \sigma^2) \propto N(\beta_0; 0, 10^2) N(\beta_1; 0, 1^2) \frac{1}{\sigma^2}$.

Subjective Bayesian regression

```
m = arm::bayesglm(chirps~temp,
                  prior.mean.for.intercept  = 0,   # E[\beta_0]
                  prior.scale.for.intercept = 10,  # SD[\beta_0]
                  prior.df.for.intercept    = Inf, # normal prior for \beta_0
                  prior.mean  = 0,                 # E[\beta_1]
                  prior.scale = 1,                 # SD[\beta_1]
                  prior.df    = Inf,               # normal prior
                  scaled = FALSE)                  # scale prior?
```

Subjective Bayesian regression

```
summary(m)


##
## Call:
## arm::bayesglm(formula = chirps ~ temp, prior.mean = 0, prior.scale = 1,
##     prior.df = Inf, prior.mean.for.intercept = 0, prior.scale.for.intercept = 10,
##     prior.df.for.intercept = Inf, scaled = FALSE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7450  -0.5795   0.0312   0.5846   1.5142
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.53636    2.99849  -0.179    0.861
## temp         0.21470    0.03738   5.743 6.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.9701008)
##
##     Null deviance: 41.993  on 14  degrees of freedom
## Residual deviance: 12.611  on 13  degrees of freedom
## AIC: 45.966
##
## Number of Fisher Scoring iterations: 10
```

Subjective vs Default

```
tmp = lm(chirps~temp) # default analysis
tmp$coefficients
```

```
## (Intercept)        temp
##  -0.6152146   0.2156787
```

```
confint(tmp)
```

```
##                  2.5 %    97.5 %
## (Intercept) -7.4081577 6.1777286
## temp         0.1310169 0.3003406
```

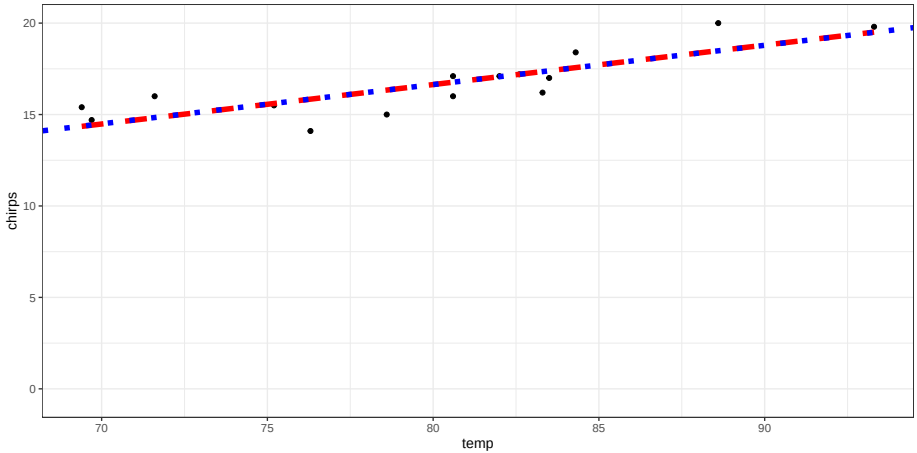
```
m$coefficients # subjective analysis
```

```
## (Intercept)        temp
##  -0.5363623   0.2146971
```

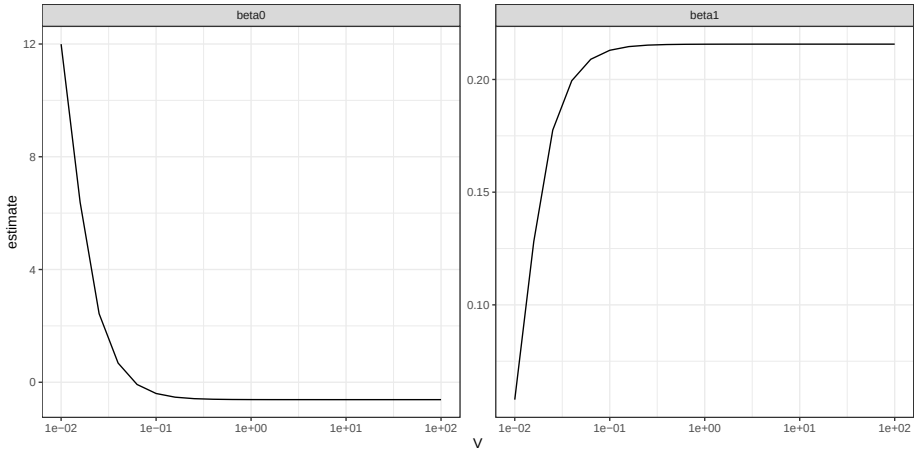
```
confint(m)
```

```
##                  2.5 %    97.5 %
## (Intercept) -6.7792735 5.5475553
## temp         0.1388709 0.2925027
```

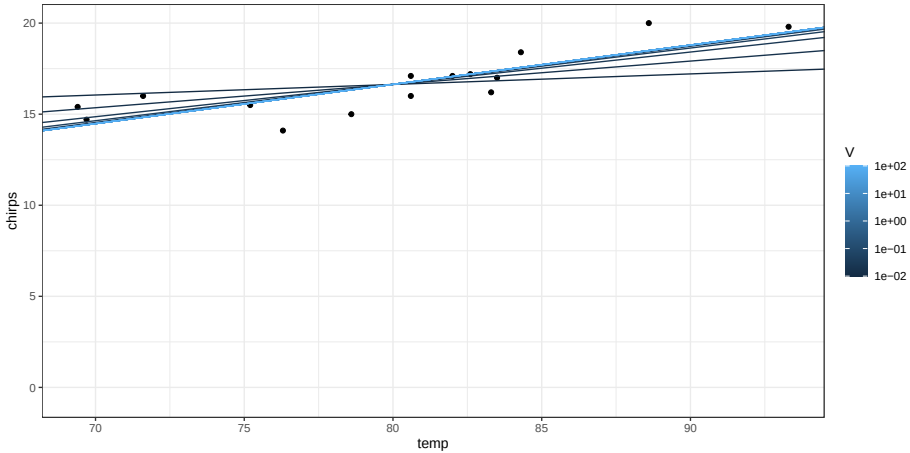
Subjective vs Default



Shrinkage (as $V[\beta_1]$ gets smaller)



Shrinkage (as $V[\beta_1]$ gets smaller)



Zellner's g-prior

Let

$$y = X\beta + \epsilon, \quad \epsilon \sim N(\sigma^2 \mathrm{I}).$$

If we choose the conjugate prior $\beta \sim N(b_0, \sigma^2 B_0)$, we still need to choose b_0 and B_0. It seems natural to set $b_0 = 0$ which will shrink the estimates for β toward zero, i.e. toward no effect. But how should we choose B_0?

One option is Zellner's g-prior where $B_0 = g[X^\top X]^{-1}$ where g is either set or learned.

Zellner's g-prior posterior

Suppose

$$y \sim N(X\beta, \sigma^2 \mathrm{I})$$

where X is $n \times p$ and you use Zellner's g-prior

$$\beta \sim N(b_0, g\sigma^2 (X'X)^{-1})$$

and assume $p(\sigma^2) \propto 1/\sigma^2$.

The posterior is then

$$\beta|\sigma^2, y \sim N\left(\frac{1}{1+g} b_0 + \frac{g}{1+g}\hat{\beta}_{MLE}, \frac{\sigma^2 g}{g+1}(X'X)^{-1}\right)$$

Setting g

In Zellner's g-prior,

$$\beta \sim N(b_0, g\sigma^2(X'X)^{-1}), p(\sigma^2) \propto 1/\sigma^2$$

we need to determine how to set g.

Here are some thoughts:

- $g \to 0$ makes posterior equal to the prior,
- $g = 1$ puts equal weight to prior and likelihood,
- $g = n$ means prior has the equivalent weight of 1 observation,
- $g \to \infty$ recovers a uniform prior,
- empirical Bayes estimate of g, $\hat{g}_{EB} = \text{argmax}_g p(y|g)$, or
- put a prior on g and perform a fully Bayesian analysis.

Marginal likelihood

The marginal likelihood under Zellner's g-prior is

$$p(y|g) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{n-1}{2}} n^{1/2}} ||y - \overline{y}||^{-(n-1)} \frac{(1+g)^{\frac{n-p-1}{2}}}{(1+g[1-R^2])^{\frac{n-1}{2}}}$$

where R^2 is the coefficient of determination.

We use the marginal likelihood as evidence in favor of the model, i.e. when comparing models those with higher marginal likelihoods should be preferred over the rest.

Why the marginal likelihood?

By Bayes' rule, we have

$$p(\theta|y, M) = p(y|\theta, M)p(\theta|M)/p(y|M)$$

Rearranging yields

$$p(y|M) = p(y|\theta, M)p(\theta|M)/p(\theta|y, M)$$

Taking logarithms yields

$$\log p(y|M) = \log(y|\theta, M) + \log p(\theta|M) - \log p(\theta|y, M)$$

To compare with other model selection criterion, multiply by -2 and plug in $\theta = \hat{\theta}_{MLE}$:

$$-2 \log p(y|M) = -2 \log(y|\hat{\theta}_{MLE}, M) + 2 \left[\log p(\hat{\theta}_{MLE}|y, M) - \log p(\hat{\theta}_{MLE}|M)\right]$$

where the penalty is the logarithm of the ratio of the posterior to the prior evaluated at the MLE.

Model selection

If β is a vector of length p, let γ be a vector with binary elements that indicate whether that component of β is non-zero, i.e. that explanatory variable is included. For example,

$$\gamma = (1, 0, 1, 1, 0, 0, 0, 1)$$

indicates that β is of length 8 and that the first, third, fourth, and eighth elements are non-zero. Then we have X_γ which indicates the design matrix that only has columns corresponding to those columns in γ that are non-zero and β_γ is the subset of β including elements of β where γ is 1.

Now, we have 2^p models M_γ of the form

$$y = X_\gamma \beta_\gamma + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 \mathrm{I})$. Two special cases are

$$\begin{aligned} \gamma_{null} &= (0, \ldots, 0) \\ \gamma_{full} &= (1, \ldots, 1) \end{aligned}$$

Model selection (cont.)

If we want to compare M_γ to M_{null} using a common g, we can use the Bayes Factor

$$BF(M_\gamma : M_{null}) = \frac{p(y|M_\gamma, g)}{p(y|M_{null}, g)} = \frac{(1+g)^{\frac{n-p_\gamma-1}{2}}}{(1+g[1-R^2_\gamma])^{\frac{n-1}{2}}}$$

Then, for any two models with a common g, we can compare these models using

$$BF(M_\gamma : M_{\gamma'}) = \frac{p(y|M_\gamma, g)}{p(y|M_{\gamma'}, g)} = \frac{p(y|M_\gamma, g)/p(y|M_{null}, g)}{p(y|M_{\gamma'}, g)/p(y|M_{null}, g)} = \frac{BF(M_\gamma : M_{null})}{BF(M_{\gamma'} : M_{null})}$$

If the base model is the null model, then the common parameters amongst the models are σ^2 and possibly an intercept α. We can place an improper prior on these parameters, typically $p(\alpha, \sigma^2) \propto 1/\sigma^2$, and not affect the Bayes Factors.

Zellner's g-prior in R

```
library(BMS)
m0 = zlm(chirps~1    , g='UIP') # g=n
m1 = zlm(chirps~scale(temp), g='UIP') # g=n
(bf = exp(m1$marg.lik-m0$marg.lik))


## [1] 438.2629


summary(m1)


## Coefficients
##              Exp.Val.   St.Dev.
## (Intercept) 16.633333        NA
## scale(temp)  1.358165 0.2839367
##
##  Log Marginal Likelihood:
## -20.07976
##  g-Prior: UIP
## Shrinkage Factor: 0.938
```

Zellner's g-prior in R

```
library(bayess)
m = BayesReg(chirps, temp, g=length(chirps)) # explanatory variables are scaled


##
##           PostMean PostStError Log10bf EvidAgaH0
## Intercept  16.6333      0.2833
## x1          1.3121      0.2743  2.6417    (****)
##
##
## Posterior Mean of Sigma2: 1.2039
## Posterior StError of Sigma2: 1.7783
```

Limiting Bayes Factors

If the base model is the null model, then

$$BF(M_\gamma : M_{null}) = \frac{(1+g)^{(n-p_\gamma-1)/2}}{(1+g[1-R^2_\gamma])^{(n-1)/2}}$$

where p_γ is the number of non-zero elements in γ, i.e. the number of explanatory variables included in the model.

- As $g \to \infty$, $BF(M_\gamma : M_{null}) \to 0$. (Lindley's Paradox)
- As $n \to \infty$, $BF(M_\gamma : M_{null}) \to \infty$.
- As $R^2_\gamma \to 1$, $BF(M_\gamma : M_{null}) \to (1+g)^{(n-p_\gamma-1)/2}$. (information paradox)

If M^* is the true model, we would like

$$BF(M^* : M_\gamma) \xrightarrow{a.s.} \infty, \quad \text{as } n \to \infty$$

for any other model M_γ. This is called model selection consistency.

Empirical Bayes

The empirical Bayes approach chooses g such that it maximizes $p(y|M_\gamma, g)$. It turns out that $g_\gamma^{EB} = \max(F_\gamma - 1, 0)$, where

$$F_\gamma = \frac{R_\gamma^2/p_\gamma}{(1-R_\gamma^2)/(n-p_\gamma-1)}.$$

Plugging this back into the expression for the Bayes Factor, we find that

$$BF^{EB}(M_\gamma : M_{null}) \to \infty$$

as $R_\gamma \to 1$ and thus the empirical Bayes approach does not suffer from either paradox. This empirical Bayes approach is model selection consistent if the true model is not the null model, but is inconsistent if it is.

Fully Bayesian

Alternatively, we can perform a fully Bayes analysis by putting a prior on g. The Zellner-Siow prior is

$$g \sim IG\left(\frac{1}{2}, \frac{n}{2}\right)$$

For this prior, we have $BF^{EB}(M_\gamma : M_{null}) \to \infty$ as $R^2_\gamma \to 1$ and thus do not suffer from any paradoxes and we have model selection consistency, i.e. $BF(M^* : M_\gamma) \xrightarrow{a.s.} \infty$ for true model M^* compared to any other model M_γ.

There are other priors for g that have these properties.