Estimating ANOVA Models with rstanarm

Jonah Gabry and Ben Goodrich

2018-04-13

Contents

Introduction

This vignette explains how to estimate ANalysis Of VAriance (ANOVA) models using the `stan_aov` function in the **rstanarm** package

The four steps of a Bayesian analysis are

1. Specify a joint distribution for the outcome(s) and all the unknowns, which typically takes the form of a marginal prior distribution for the unknowns multiplied by a likelihood for the outcome(s) conditional on the unknowns. This joint distribution is proportional to a posterior distribution of the unknowns conditional on the observed data
2. Draw from posterior distribution using Markov Chain Monte Carlo (MCMC).
3. Evaluate how well the model fits the data and possibly revise the model.
4. Draw from the posterior predictive distribution of the outcome(s) given interesting values of the predictors in order to visualize how a manipulation of a predictor affects (a function of) the outcome(s).

Steps 3 and 4 are covered in more depth by the vignette entitled "How to Use the **rstanarm** Package" (rstanarm.html). This vignette focuses on Step 1 when the likelihood is the product of independent normal distributions. We also demonstrate that Step 2 is not entirely automatic because it is sometimes necessary to specify some additional tuning parameters in order to obtain optimally efficient results.

Likelihood

The likelihood for one observation under a linear model can be written as a conditionally normal PDF

$$\frac{1}{\sigma_\epsilon \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma_\epsilon}\right)^2},$$

where $\mu = \alpha + \mathbf{x}^\top \boldsymbol{\beta}$ is a linear predictor and σ_ϵ is the standard deviation of the error in predicting the outcome, y. The likelihood of the entire sample is the product of N individual likelihood contributions.

An ANOVA model can be considered a special case of the above linear regression model where each of the K predictors in $\mathbf{x}$ is a dummy variable indicating membership in a group. An equivalent linear predictor can be written as $\mu_j = \alpha + \alpha_j$, which expresses the conditional expectation of the outcome in the j-th group as the sum of a common mean, α, and a group-specfic deviation from the common mean, α_j.

Priors

If we view the ANOVA model as a special case of a linear regression model with only dummy variables as predictors, then the model could be estimated using the prior specification in the `stan_lm` function. In fact, this is exactly how the `stan_aov` function is coded. These functions require the user to specify a value for the prior location (by default the mode) of the R^2, the proportion of variance in the outcome attributable to the predictors under a linear model. This prior specification is appealing in an ANOVA context because of the fundamental identity

$$SS_{\text{total}} = SS_{\text{model}} + SS_{\text{error}},$$

where SS stands for sum-of-squares. If we normalize this identity, we obtain the tautology $1 = R^2 + \left(1 - R^2\right)$ but it is reasonable to expect a researcher to have a plausible guess for R^2 before conducting an ANOVA. See the vignette (lm.html) for the `stan_lm` function (regularized linear models) for more information on this approach.

If we view the ANOVA model as a difference of means, then the model could be estimated using the prior specification in the `stan_lmer` function. In the syntax popularized by the **lme4** package, `y ~ 1 + (1|group)` represents a likelihood where $\mu_j = \alpha + \alpha_j$ and α_j is normally distributed across the J groups with mean zero and some unknown standard deviation. The `stan_lmer` function specifies that this standard deviation has a Gamma prior with, by default, both its shape and scale parameters equal to 1, which is just an standard exponential distribution. However, the shape and scale parameters can be specified as other

positive values. This approach also requires specifying a prior distribution on the standard deviation of the errors that is independent of the prior distribution for each α_j. See the vignette (glmer.html) for the `stan_glmer` function (**lme4**-style models using **rstanarm**) for more information on this approach.

Example

We will utilize an example from the **HSAUR3** package by Brian S. Everitt and Torsten Hothorn, which is used in their 2014 book *A Handbook of Statistical Analyses Using R (3rd Edition)* (Chapman & Hall / CRC). This book is frequentist in nature and we will show how to obtain the corresponding Bayesian results.

The model in section 4.3.1 analyzes an experiment where rats were subjected to different diets in order to see how much weight they gained. The experimental factors were whether their diet had low or high protein and whether the protein was derived from beef or cereal. Before seeing the data, one might expect that a moderate proportion of the variance in weight gain might be attributed to protein (source) in the diet. The frequentist ANOVA estimates can be obtained:

```
data("weightgain", package = "HSAUR3")
coef(aov(weightgain ~ source * type, data = weightgain))
```

```
        (Intercept)         sourceCereal              typeLow
              100.0                -14.1                -20.8
sourceCereal:typeLow
               18.8
```

To obtain Bayesian estimates we can prepend `stan_` to `aov` and specify the prior location of the R^2 as well as optionally the number of cores that the computer is allowed to utilize:

```
library(rstanarm)
post1 <- stan_aov (../reference/stan_lm.html)(weightgain ~ source * type, data = weightgain,
                   prior = R2 (../reference/priors.html)(location = 0.5), adapt_delta = 0.999,
                   chains = CHAINS, cores = CORES, seed = SEED)
post1
```

```
stan_aov
 family:       gaussian [identity]
 formula:      weightgain ~ source * type
 observations: 40
------
                    Median MAD_SD
(Intercept)          98.9    4.5
sourceCereal        -12.9    6.4
typeLow             -18.6    6.4
sourceCereal:typeLow  16.9    9.0
sigma                 14.8    1.8
log-fit_ratio          0.0    0.1
R2                     0.2    0.1

Sample avg. posterior predictive distribution of y:
         Median MAD_SD
mean_PPD 87.3    3.2

ANOVA-like table:
                   Median MAD_SD
Mean Sq source      565.0  452.8
Mean Sq type        979.5  594.8
Mean Sq source:type 713.5  710.2

------
For info on the priors used see help('prior_summary.stanreg').
```

Here we have specified `adapt_delta = 0.999` to decrease the stepsize and largely prevent divergent transitions. See the Troubleshooting section in the main rstanarm vignette (rstanarm.html) for more details about `adapt_delta`. Also, our prior guess that $R^2 = 0.5$ was overly optimistic. However, the frequentist estimates presumably overfit the data even more.

Alternatively, we could prepend `stan_` to `lmer` and specify the corresponding priors

```
post2 <- stan_lmer (../reference/stan_glmer.html)(weightgain ~ 1 + (1|source) + (1|type) + (1|sourc
                   data = weightgain, prior_intercept = cauchy (../reference/priors.html)(),
                   prior_covariance = decov (../reference/priors.html)(shape = 2, scale = 2),
                   adapt_delta = 0.999, chains = CHAINS, cores = CORES,
                   seed = SEED)
```

Comparing these two models using the `loo` function in the **loo** package reveals a negligible preference for the first approach that is almost entirely due to its having a smaller number of effective parameters as a result of the more regularizing priors. However, the difference is so small that it may seem advantageous to present the second results which are more in line with a mainstream Bayesian approach to an ANOVA model.

Conclusion

This vignette has compared and contrasted two approaches to estimating an ANOVA model with Bayesian techniques using the **rstanarm** package. They both have the same likelihood, so the (small in this case) differences in the results are attributable to differences in the priors.

The `stan_aov` approach just calls `stan_lm` and thus only requires a prior location on the R^2 of the linear model. This seems rather easy to do in the context of an ANOVA decomposition of the total sum-of-squares in the outcome into model sum-of-squares and residual sum-of-squares.

The `stan_lmer` approach just calls `stan_glm` but specifies a normal prior with mean zero for the deviations from α across groups. This is more in line with what most Bayesians would do naturally — particularly if the factors were considered "random" — but also requires a prior for α, σ, and the standard deviation of the normal prior on the group-level intercepts. The `stan_lmer` approach is very flexible and might be more appropriate for more complicated experimental designs.