Focused model comparison with bootstrapping and alternative loss functions

Christopher Jackson
chris.jackson@mrc-bsu.cam.ac.uk

Abstract

This vignette briefly illustrates a bootstrap procedure for focused model comparison which allows very general loss functions, though is based on different asymptotic assumptions from standard focused model comparison. This procedure is experimental and has not been studied in detail.

Keywords: models,bootstrap.

Take the low birth weight example, described in detail in the main FIC package vignette. We estimate the probability of low birth weight for smokers using a logistic regression, and compare 64 subsets of the wide model with different combinations of covariates. Focused model comparison statistics are computed, as before.

```
library(fic)
wide.glm <- glm(low ~ lwtkg + age + smoke + ht + ui + smokeage + smokeui,
                data=birthwt, family=binomial)
vals.smoke <- c(1, 58.24, 22.95, 1, 0, 0, 22.95, 0)
X <- vals.smoke
inds0 <- c(1,1,0,0,0,0,0,0)
combs <- all_inds(wide.glm, inds0)
ficres <- fic(wide=wide.glm, inds=combs, inds0=inds0,
              focus=prob_logistic, X=X)
```

These focused model comparison statistics are computed using formulae derived under an asymptotic framework where we assume that the data are generated from a model with parameters which depart from the parameters of the narrow model by an amount $\boldsymbol{\delta}/\sqrt{n}$ that depends on the sample size n.

An alternative approach might be to assume that the data were generated under a wide model that does not depend on the sample size. Then, to compute the mean square error of the estimate of a focus quantity $\mu(\boldsymbol{\theta}, \boldsymbol{\gamma})$ under a submodel with estimated parameters $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_S, \boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}_S$, we generate a large number B of alternative parameter estimates $(\boldsymbol{\theta}^{(r)}, \boldsymbol{\gamma}^{(r)})$ from the multivariate normal sampling distribution defined by the maximum likelihood estimates and covariance matrix from the wide model.

The mean square error of the focus is then estimated as

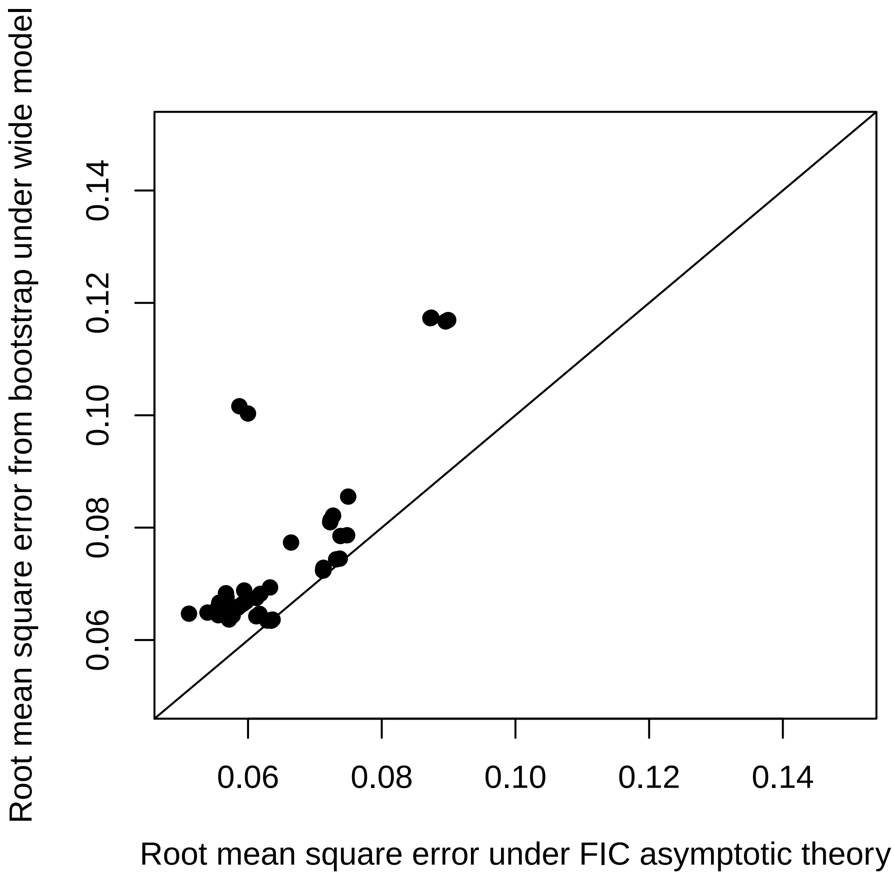
$$\frac{1}{B}\sum_{r=1}^{B}\left(\mu(\boldsymbol{\theta}^{(r)}, \boldsymbol{\gamma}^{(r)}) - \mu(\hat{\boldsymbol{\theta}}_S, \hat{\boldsymbol{\gamma}}_S)\right)^2$$

This is implemented in the `fic` function by supplying a `B` argument giving the number of bootstrap samples.

```
ficboot_mse <- fic(wide=wide.glm, inds=combs, inds0=inds0,
                   focus=prob_logistic, X=X, B=1000)
```

The root mean square errors for each of the 64 models are compared here to those estimated under the standard framework with a sample-size dependent true model. The model preference is similar between the two methods, with a small handful of submodels where the methods give different estimates of error.

```
plot(ficres$rmse, ficboot_mse$loss, xlim=c(0.05,0.15),
     ylim=c(0.05,0.15), pch=19,
     xlab = "Root mean square error under FIC asymptotic theory",
     ylab = "Root mean square error from bootstrap under wide model")
abline(a=0, b=1)
```



This framework allows alternative loss functions, for example, the absolute error loss:

$$\frac{1}{B}\sum_{r=1}^{B}\left|\mu(\boldsymbol{\theta}^{(r)},\boldsymbol{\gamma}^{(r)}) - \mu(\hat{\boldsymbol{\theta}}_S,\hat{\boldsymbol{\gamma}}_S)\right|$$

Alternative loss functions can be supplied to `fic` as functions of the (scalar) submodel estimate and the vector of bootstrapped wide model estimates, illustrated here for the absolute error:

```
loss_abserror <- function(sub, wide){
    mean(abs(sub - wide))
}
ficboot_abs <- fic(wide=wide.glm, inds=combs, inds0=inds0,
                   focus=prob_logistic, X=X, B=1000, loss=loss_abserror)
```

While the bootstrap approach is computationally convenient, the relative properties of the two different asymptotic frameworks have not been studied.